GMR
The Original

Customer Behavior Analysis Using Cascade Adaptive Feature Reconstruction And Encoding In E-Commerce Datasets

Mr. K.Dhiyaneshwaran[1], Dr. P. Sharmila[2]

[1]Research Scholar, Navarasam Arts and Science College for Women, Affiliated to Bharathiar University, Arachalur - 638101, Tamil Nadu, India. Email: dhiyaneskailas@gmail.com, ORCID: 0009-0000-5268-0923
[2]Professor and Dean, School of Computing Science, KPR College of Arts Science and Research, Affiliated to Bharathiar University, Arasur, Coimbatore-641407, Tamil Nadu, India. ORCID: 0009-0000-6466-3549

Abstract
E-commerce is the sale and purchase of goods or services through the internet, enabled by online platforms for user interaction, product reviews and transactions. Under the digital platform, customer feedback is essential in enhancing product quality, user experience and business decision-making. This paper suggests a new two-stage approach for enhancing sentiment analysis and review quality prediction in e-commerce based on Amazon product reviews. In stage one, the Hierarchical Intra-Session Behavior Adaptive Reconstructions (HBA-REC) approach preprocesses user reviews by dividing these into micro and meta sessions to identify the short- and long-term behavior patterns. HBA-REC improves data reliability by maintaining rare events, recovering truncated reviews and incorporating contextual sentiment, time and interaction dynamics. In stage two, the Cascade Adaptive Feature Reconstruction and Encoding (CAFRC) system conducts adaptive feature selection by multi-layer analysis through L1 regularization, autoencoder reconstruction loss and integrated gradients. CAFRC masks unstable or redundant features while retaining features with high semantic and predictive importance. Experimental comparisons demonstrate that HBA-REC outperforms conventional and deep learning-based preprocessing methods significantly with lower Mean Absolute Error (MAE) (0.26), Mean Squared Error (MSE) (0.22) and improved R^2 score (0.88). Likewise, CAFRC performs the highest classification accuracy (92.89%) compared to feature selection methods and outperforms Chi-Square, Principal Component Analysis (PCA) and L1-based methods. The combined pipeline produces a behavior-enriched high-quality feature set that enables strong and interpretable sentiment prediction models. Results indicate improved model performance, reduced overfitting and improved generalizability on e-commerce tasks. Together, HBA-REC and CAFRC constitute an integrated preprocessing and feature optimization pipeline that greatly enhances the efficiency and reliability of e-commerce analytics.

Keywords: Behavior-Aware Preprocessing, Feature Selection, Amazon Product Reviews, Sentiment Analysis.

I. Introduction

E-commerce platforms have seen a quick rise in data volume because more users are engaging across various devices and touchpoints. This development necessitates better techniques of data handling, modeling behavior and predicting. It is the information between details or the chronological correlation of

details which is crucial to the realization of personalized e-commerce endeavors that is normally overshadowed in a conventional data processing and modeling of sessions [1], [2]. The right selection of features and the hybridization is critical in executing the extraction of useful information out of complicated and noisy data. This, in turn, boosts up the Machine Learning (ML) procedures [3].
Recently, reports demonstrate the importance of cleaning the e-commerce data prior to working on it through predictive analytics [4]. Combining the feature selection and explainable AI (XAI) in describing the product development in terms of predictive behavior of the customers in relation to the customer churn has also been successful [5]. Inefficiently designed preprocessing pipeline goes on other way and bring the lower levels of prediction accuracy, decrease customer satisfaction [6] and dropout rates [7].
In addition, the idea of reduction in dimension and feature selection with the use of filters has been considerable in carrying out the predictions of specific numbers in e-commerce. An example of such jobs is to analyze poverty [8], find fraud [9] and analyze customer attrition [10]. The necessity to have more sets which contain data with temporal data as well as contextual data is beyond the reach of predictive models to purchase behavior [11] and those that increase the awareness to time sensitive suggestions [12].
Despite these new developments, some of the existing frameworks have slowdowns in identifying uncommon behavior patterns and correlation among different features. This type of problem is especially noticeable in the case of separating such data into micro and meta sessions. Consequently, recommendation or forecasting systems are not adaptable to the changes in environment [13], [14]. Besides, another study that uses sentiment data in business decision research provides evidence that there has slowly been an increment of clear preprocessing techniques [15].
To overcome these, a hybrid behavioral and feature reconstruction system is put forward. This proposal is a union between Hierarchical Intra-session Behavior Adaptive Reconstructions (HBA-REC) and the Cascade Adaptive Feature Reconstruction and Encoding (CAFRC). This integrated approach allows for the reconstruction of user behavior across fragmented sessions while simplifying and optimizing features using domain-specific strategies. The outcome is a dataset that is consistent over time, enriched with behavior data and reduced in noise, designed specifically for e-commerce machine learning applications.

Contribution: A dual-layered mechanism reconstructs behavior to keep rare and out-of-order interactions across sessions. A cascaded process refines features by filtering out low-quality and redundant ones while improving task-specific data quality. Empirical validation using e-commerce datasets shows better accuracy, less overfitting and improved interpretability.

Organization: The rest of the paper is organized in this way: Section 2 presents related work. Section 3 details the proposed framework. Section 4 discusses experimental results. Section 5 concludes with implications and future research directions.

II. Background Study

Mirdan et al. (2025) [16] conducted a detailed sentiment analysis on Twitter data related to Amazon. The authors used ML classifiers to assess consumer perception and platform performance. The study highlighted the strong link between customer sentiment and e-commerce brand loyalty. The authors also looked at the model performance across various algorithms and stressed the effectiveness of Natural Language Processing (NLP) pipelines for real-time sentiment tracking. This work was supported by using the external behavioral signals to improve recommendation systems.
Prabhakaran and Nedunchelian (2023) [17] proposed a feature selection method based on Oppositional Cat Swarm Optimization (OCSO) to improve fraud detection accuracy in e-commerce transactions. The algorithm boosted classifier performance by picking only the most relevant features. This reduced computational overhead and false positives. The hybrid approach showed the strong results on benchmark datasets, especially in effectively identifying the fraudulent credit card activities.

Farsi and Chowdhury (2025) [18] introduced EcomFraudEX, a new explainable ML framework for detecting fraud from both victims and perpetrators. The authors used interpretable ML models to classify the incidents with high accuracy while ensuring transparency in decision-making. The authors aimed the approach at the often neglected victim side of fraud model and had the work confirmed by using actual transaction data.

Pustokhina et al. (2021) [19] proposed a dynamic churn prediction model enhancing it by using text analysis as well as optimization algorithms so that these retained the customers in an e-commerce environment. In this system, reviews were examined on customers and the behavioral data using ML and optimization algorithms to enhance the predictability of same. The research emphasized that more and more flexibility of churn modeling was required and that textual customer data was presented as significant business intelligence.

Gupta et al. (2023) [20] presented a composite sentiment analysis code that used countless ML approaches to categorize user views in online sale appraisal. The authors have used text preprocessing, polarity score and classifier ensembles to enhance the accuracy of classification. The model had helped in getting a better idea of feelings and preferences of the users which had been important in improving suggestions and personalization systems.

Table 1: Comparison Table on Product Reviews and Customer Feedback

Authors (Year)	**Objective**	**Methodology**	**Data/Language Used**	**Limitations**
Savci & Das (2023) [21]	Predict customer interests via sentiment analysis	Comparative sentiment classification	Arabic, English, Turkish	Language-wise differences influence prediction accuracy; English performs best
Gupta et al. (2025) [22]	Detect credit card fraud effectively	Robust feature selection with Stacking Ensemble	Financial transaction datasets	Improved fraud detection accuracy with reduced false positives
Daoud & Kammoun (2024) [23]	Forecast e-commerce adoption in SMEs	ML on behavioral drivers	Small and Medium-sized Enterprises (SME) behavioral/economic data	Identified key predictors of adoption; high model interpretability
Vijayaraga van et al. (2024) [24]	Perform sustainable sentiment analysis for smart cities	Weighted parallel hybrid Deep Learning (DL)	E-commerce review data	Enhanced accuracy and scalability for smart city recommendation apps
Bagwari et al. (2022) [25]	Develop business-decision content recommendation model	Content-Based Image Retrieval (CBIR-DSS)	Visual and content data	Decision-oriented recommendations improved conversion potential

In a recent paper, Esmeli and Gokce (2025) [26] examined the behavior of consumers when the objects were placed in the cart using the Explainable Artificial Intelligence (XAI) method. In the study, the authors discovered the reasons and time delays of cart abandonment and purchases. The emphasis on explainability provided the real-life implications to enhance checkout procedures and customize methods in a web-based commerce system.

In the study of logistics performance and economic indicators, Jomthanachai et al. (2022) [27] explored the regression-based ML to determine these features selection. The integration approach enhanced the model correctness by recognizing crucial attributes. This displayed its significance as a supply chain analysis and logistic forecasting in e-commerce.

Alotaibi (2023) [28] proposed a methodology based on ML to overcome this by extracting and classifying the customer opinion about the review of customers on social media. The system involved the text analysis and the evaluation of customer satisfaction levels. The research was found to be more precise in segregating the various emotional responses and this assisted business houses to cope with e-commerce reputation.

Explainable AI (XAI) methods were also adopted by Alotaibi et al. (2025) [29] to label the phishing attempts at websites associated with secure Internet of Things (IoT)-environments and cloud-based cyber-physical systems. The architecture enhanced cyberspace intelligence as well as bolstering cybersecurity that was crucial for safeguarding e-commerce platforms against the risks of cyber-attacks.

To have a better understanding of the customer behavior, Subramanian and Prabha (2022) [30] suggested an ensemble approach for selecting the variables in Naive Bayes classifiers. The approach enhanced feature selection and classification in the consumer analytics dimension and it had been applicable in conducting targeted marketing and behavior segmentation functions.

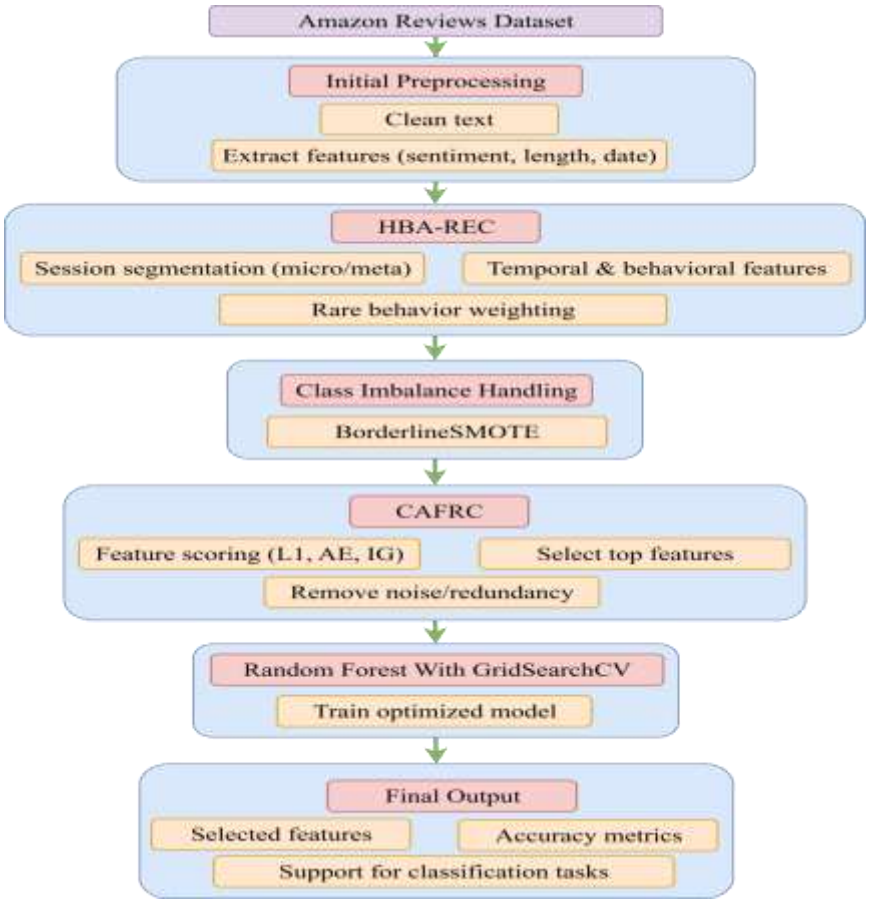
2.1 Problem Identification

Despite the increasing use of ML in e-commerce, major challenges still exist. These challenges include modeling complex user behavior, detecting fraud transparently and handling large data sets in multiple languages. Often, current methods have not provided a clear framework that connects behavior reconstruction, explainability and feature optimization across different areas. There is a strong need for systems that integrate sentiment analysis, fraud signals and behavioral insights while ensuring interpretability, data quality and flexibility in various e-commerce environments.

III. Materials and Methods

With the age of e-commerce, customer reviews constitute a rich source of user feedback that impacts product perception, user interaction and business strategy. This section details the comprehensive methodology adopted for preprocessing and feature selection in the context of sentiment analysis and product quality prediction using Amazon review data. The raw dataset, sourced from Kaggle includes over 34,000 reviews with the attributes such as review text, rating, brand and date. The proposed pipeline consists of two core components: Hierarchical Intra-Session Behavior Adaptive Reconstructions (HBA-REC) for semantically and behaviorally enriched preprocessing and Cascade Adaptive Feature Reconstruction & Encoding (CAFRC) for robust feature selection. Together, these methods ensure the transformation of noisy, high-dimensional data into a structured, balanced and model-ready form suitable for high-performance machine learning tasks.

Figure 1: Overall Work Flow Architecture



This figure 1 represents a pipeline for classifying Amazon reviews. The pipeline starts with preprocessing the textual data and engineering particular features like sentiment and word count. After that, the HBA-REC module segments sessions and computes the temporal and behavioral patterns and rare behaviors. Once the feature engineering process is complete, the class imbalance is addressed during the sampling process using BorderlineSMOTE to create the synthetic samples. Then, the CAFRC module scores the features and selects the most informative to remove noise and redundancy. Finally, the pipeline finishes with an optimized Random Forest model trained using GridSearchCV and outputs the selected features, accuracy metrics and mapped outputs for other classification tasks.

3.1 Dataset

https://www.kaggle.com/datasets/yasserh/amazon-product-reviews-dataset. From the Kaggle dataset "Amazon Product Reviews" by Yasser H. includes more than 34,000 customer reviews. It has features such as review text, rating, product title, brand and date. This dataset is commonly used for sentiment analysis, predicting product ratings and modeling behavior in e-commerce analytics.

3.2 Preprocessing using Hierarchical Intra-Session Behavior Adaptive Reconstructions (HBA-REC)

The Hierarchical Intra-Session Behavior Adaptive Reconstructions (HBA-REC) is a comprehensive preprocessing algorithm developed to transform noisy, unstructured Amazon product review data into structured, semantically meaningful and model-ready input for sentiment analysis and product quality prediction tasks. Unlike traditional preprocessing pipelines, HBA-REC uses both semantic encoding and temporal behavioral features to capture the true intent, context and user review patterns.

The preprocessing starts with scrolling and reviewing the data stuck up, dealing with the blanked out values and changing the text to small letter. First, sentiment polarity, subjectivity, length of the review and month written are performed. To discover secret trends in the distribution of ratings, dynamic of the sentiments and time-related changes, elevated Exploratory Data Analysis (EDA) is conducted.

The most relevant point in HBA-REC algorithm is the ability to clean the text in a hierarchical manner to improve the quality and contextual material of review data. It starts by eliminating the vicious data like URLs, non alphabetic characters and uncommon stop words that have not added the useful information. After cleaning, sentiment polarity is again calculated to be apt in terms of the text that has been refined. Since the algorithm represents the temporal features of the behavior such as review day, age and weekday are been extracted to better model the user engagement patterns and time-sensitive sentiment patterns.

The reviews are then semantically encoded in terms of using the Sentence Transformer model wherein the unstructured text form of reviews is converted into dense and fixed length of the vectors. Given a review text R_i, the Sentence Transformer f maps that text into a high dimensional semantic embedding space:

$E_i = f(R_i) \in R^d$ ---------- (1)

In equation (1), E_i is the embedding vector of the i^{th} review and d is embedding dimension. These embeddings model contextual relationships and semantics that go beyond the keywords frequency and is able to model downstream accurately.

Since sectors are now characterized due to high dimensionality of embeddings, Principal Component Analysis (PCA) is referred from (Sachin, D. (2015) [31]) has been used to cut down the computational burden due to the many dimensions involved, without losing trends dominating behavior. PCA projects the data into a lower-dimensional subspace by selecting the top k components that maximize variance:

$Z = E \cdot W_k$ ---------- (2)

The equation (2) is categorical features such as brand, category, color and manufacturer are encoded using target encoding, which replaces each category with the mean of the target variable y (e.g., product rating or sentiment class) for that category. For a given categorical feature C, the target encoded value for category c_j is computed as:

$TE(c_j) = \frac{1}{|S_{c_j}|} \sum_{i \in S_{c_j}} y_i$ ---------- (3)

Equation (3) S_{c_j} is the set of all samples with category c_j and y_i is the target label of the i^{th} sample. This allows categorical variables to retain the influence on target distribution in a continuous form suitable for model training. The numerical features (e.g., review length, sentiment polarity, age of review) are normalized using MinMax Scaling to map each value into the [0, 1] range. For a feature x, the scaled value x′ is computed as:

$x' = \frac{x - min\,(x)}{max(x) - min\,(x)}$ ---------- (4)

This equation (4) ensures that all input features contribute proportionately to the learning process, avoiding dominance by features with larger magnitudes. To address the issue of class imbalance, particularly in skewed review ratings or sentiment labels, the Borderline Synthetic Minority Oversampling Technique (BorderlineSMOTE) is used. This technique focuses on samples near the decision boundary,

which are more critical for classification. For a minority class sample x_i, new synthetic instances $\tilde{x}$ are generated using linear interpolation with one of its k-nearest neighbor's x_{zi} equation (5):

$\tilde{x} = x_i + \lambda(x_{zi} - x_i)$ ---------- (5)

This process creates new data points that reinforce the minority class near decision boundaries, improving class separability and model sensitivity without causing overfitting in less informative regions.

Existing sentiment analysis methods lack from several critical limitations. Traditional models often rely on shallow text representations such as TF-IDF, which fail to capture semantic context. The HBA-REC framework addresses this by using Sentence Transformer embeddings that encode deep contextual meaning. Moreover, while earlier approaches neglect comprehensive text cleaning, HBA-REC removes URLs, non-alphabetic tokens and stopwords, followed by recalculating sentiment features to ensure input clarity. Temporal behavior, often ignored in previous works is modeled through features like review day, age and weekday. To handle metadata appropriately, categorical variables are target encoded, and numerical features are normalized using MinMaxScaler is referred from (Deepa, B., & Ramesh, K. (2022) [32]). Finally, HBA-REC resolves class imbalance through BorderlineSMOTE, which creates synthetic samples near decision boundaries, enhancing model sensitivity to underrepresented classes.

Figure 2: HBA-REC Architecture

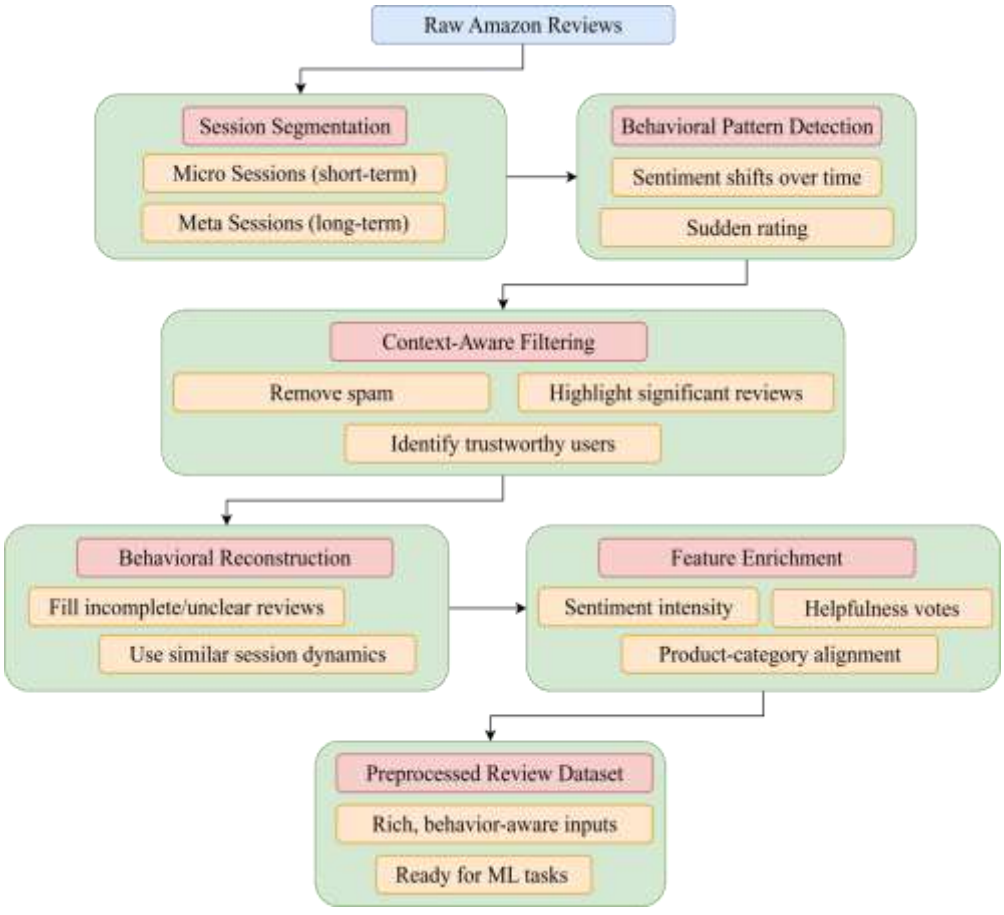


Figure 2 provides a complete preprocessing pipeline for Amazon reviews. It starts with seasoning and behavioral change detection such as the changes in sentiment and abrupt ratings. Next, context aware filtering is applied, removing spam, estimating the reviews of significance and marking these with trustworthy users. Lastly, incomplete reviews are reconstructed by labeling these with the features such as sentiment intensity and helpfulness votes, exhausted but now; a clean dataset is present reflecting the real user behavior, ready for downstream ML tasks.

3.3 Feature Selection using Cascade Adaptive Feature Reconstruction & Encoding (CAFRC)

CAFRC (Cascade Adaptive Feature Reconstruction and Encoding), a multi-criteria feature selection framework intended to increase the reliability, interpretability and performance of machine learning models in the applications such as sentiment analysis and review quality prediction. Feature selection is an important step, particularly in high-dimensional datasets where keeping only the most informative and stable attributes can greatly impact downstream learning results. The CAFRC strategy smartly combines statistical weighting, reconstruction quality and gradient-based sensitivity into a single scoring framework that extracts the most informative features with minimal redundancy and overfitting.

Existing feature selection literature has mostly depended on one-dimensional metrics like correlation coefficients, variance cut-offs or univariate statistical tests. These are useful in certain contexts but tend to neglect the features interaction with each other and the combined contribution to model performance. Moreover, most of the classic scaling methods like MinMaxScaler are very sensitive to outliers, skewing feature distribution and misdirecting the selection process. Methods like PCA and Chi-Square, while common, have a tendency to deprive the features of interpretability or neglect structural reconstruction significance.

To ensure stable model training, the input feature set X is scaled using RobustScaler, which centers the data on median and scales in according to the Interquartile Range (IQR). This approach minimizes the influence of outliers while preserving the inherent distribution of data, making it more robust compared to the standard normalization techniques.

A neural network with L1 regularization is referred from is trained on the scaled feature set to encourage sparsity in the learned weights. The regularization term penalizes large weights, effectively zeroing out less important features. The L1 score for each feature is computed as the normalized absolute value of its weight:

$$L1\_Score_j = \frac{|\omega_j|}{\sum_{k=1}^{d}|\omega_k|} \text{ ---------- (6)}$$

Equation (6) ω_j is the weight for feature j and d is the total number of features. These scores reflect the statistical importance of features in predicting the target.

An autoencoder, a type of neural network is trained to compress and reconstruct the input feature set. It learns to retain the most significant patterns necessary for accurate reconstruction. The error of reconstruction of the feature is obtained as:

$$AE\_Score_j = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2 \text{ ---------- (7)}$$

In equation (7), the original values are x_{ij} and reconstructed values $\hat{x}_{ij}$of the jth feature of the sample i and n is the number of samples. The feature which score better in term of AEs are said to be more informative because the model has problem in representation and therefore, these are relevant in the structure.

As a way of determining the functional contribution of each feature, Integrated Gradients (IG) is calculated with TensorFlow GradientTape. IG is used to quantify the effect changes in each feature of an input have had on the prediction of a model by summing the gradient of a path from a baseline input value established by forces to the real input value. IG of a feature j is calculated by:

$$IG_j = (x_j - x'_j) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_j} d\alpha \text{ ---------- (8)}$$

As per equation (8), x is the input, x' is a total baseline and F is where the model is output. This is a gradient-based measure of interpretability which measures the sensitivity of the prediction to each input.
In order to obtain exhaustive feature relevance measure, the L1 regularization (L1_Score), autoencoder reconstruction error (AE_Score) and Integrated Gradients (IG_Score) are averaged:

$CompositeScore_j = \frac{1}{3}(L1_{Score_j} + AE\_Score_j + IG\_Score_j)$ ---------- (9)

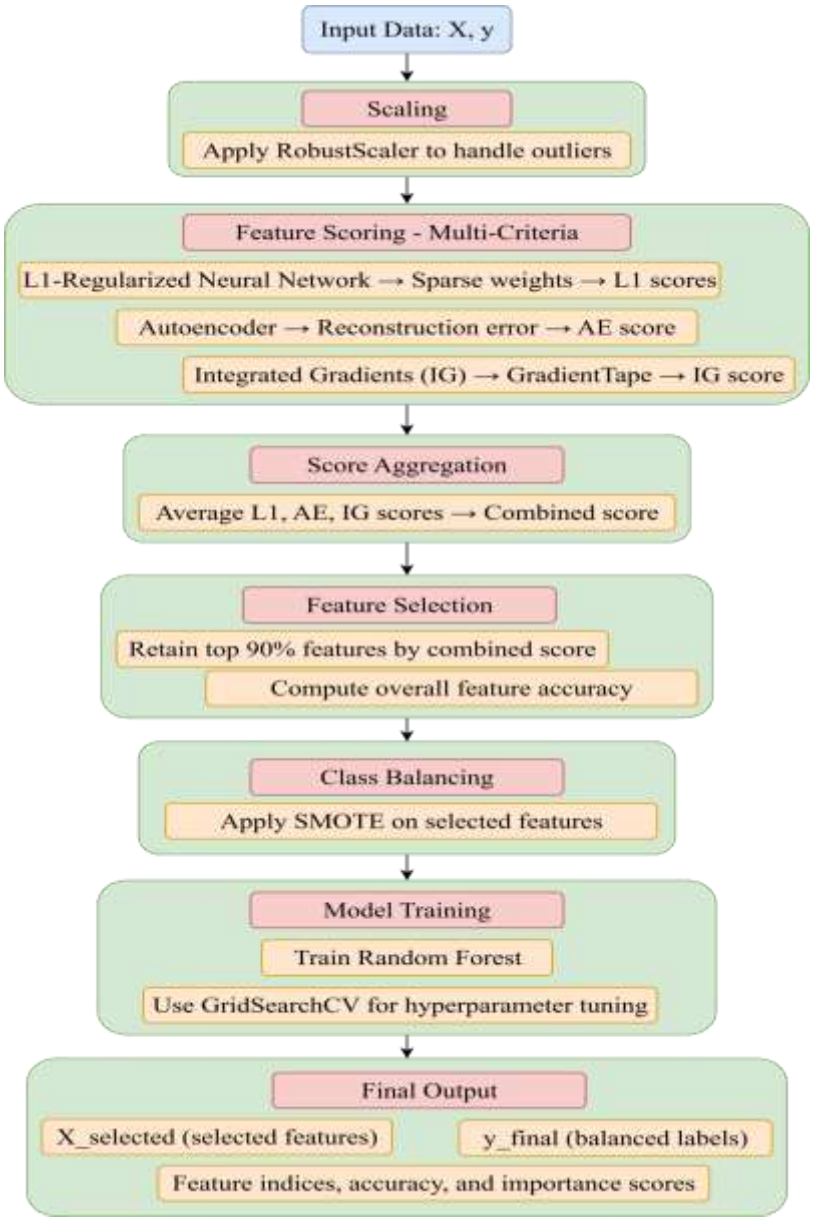
The equation (9) composite score combines statistical sparsity, structural reconstruction difficulty and functional sensitivity to yield a strong and multifaceted measure of the significance of any feature to downstream modeling.

Features are ranked top to bottom after scores are calculated regarding the composite feature importance. The index-based thresholding is applied to identify and keep 90 percent of the most informative features in order to minimize the dimensionality. The resulting selection method is capable of retaining most of the information that predicts something and subsequently trimming features that are highly irrelevant or have limited influences, thus making the model more efficient.
In order to evaluate the value of chosen features, feature accuracy measured is the ratio of the composite scores cumulative of the chosen features to the sum of the total composite scores:

ss$Feature\ Accuracy = \frac{\sum_{j \in Selected} CompositeScore_j}{\sum_{j=1}^{d} CompositeScore_j}$ ---------- (10)

The cumulative number of features is celebrated in equation (10) d. The measure establishes the extent accuracy remains intact and it serves as an objective measure of the effect of feature reduction or model fidelity.
The selected feature set is processed through the resampling procedure involving the usage of SMOTE (Synthetic Minority Over-sampling Technique) based on the suggestion that makes up Sağlam & Cengiz (2022) [33] to offset the issue of the unbalanced classes. Using SMOTE, synthetics is made on the minor classes by interpolating between examples to avoid replication in the process of adding minority representations. This assists in offering a harmonized dispersion of the classes, escalation of the classifier and rejection of the underlying bias amidst the classes that are the majorities of labels.
Thereafter, a Random Forest classifier is trained using the balanced data because it is robust and interpretable. In order to work in the most successful way, the GridSearchCV is conducted, the aim of which is to search with the help of grid of the values. Cross-validation is then done to ensure stability and then the validations accuracy selects the most accurate model. Finally, the model performance on a held-out test set in terms of generalization is identified.
The final pipeline stage combines all the obtained results that are required to implement the model and interpret. The function returns the refined dataset comprising the selected features $(X_{selected})$, the corresponding final target labels(y_{final}), the indices of selected features, the computed feature accuracy and the composite importance scores for all features. To enhance interpretability, the function also prints summary statistics such as the number of features selected and the cumulative importance and visualizes the feature importance distribution through a bar chart. This chart highlights the top-ranked features, providing clear insights into the relative contribution to the model's predictive capability.

Figure 3: CAFRC Architecture



This figure 3 presents a pipeline for feature selection and model building. First, the input is scaled for outliers. Next, features are scored using L1-regularized neural networks, autoencoders and Integrated Gradients. The scores are averaged to rank features and the implementation keeps the top 90% of features and then applies Synthetic Minority Over-sampling Technique (SMOTE) for class balancing. A Random Forest model is built with hyperparameter tuning, producing the selected features, the balanced labels and accuracy reports.

Algorithm 1: HBA-REC and CAFRC

```
BEGIN
INPUT:
  - Raw dataset: "Amazon Product Reviews.csv"
1. IMPORT LIBRARIES
  Import pandas, numpy, matplotlib, re, nltk, textblob
  Import sklearn, tensorflow, imblearn, category_encoders
  Import SentenceTransformer from sentence_transformers
2. LOAD DATA AND CONFIGURE PLOTS
  Set plot font and DPI
  Read dataset → df
  Display df.info(), df.shape, df.describe(), df.isnull().sum()
  Plot rating distribution
OUTPUT:
  - Cleaned and loaded DataFrame: df
3. INITIAL PREPROCESSING
  Drop rows with nulls in ['reviews.text', 'reviews.rating', 'reviews.date', 'dateAdded']
  Lowercase 'reviews.text' → df['cleaned_review']
  Compute polarity and subjectivity → df['polarity'], df['subjectivity']
  Compute review length → df['review_len']
  Convert 'reviews.date' to datetime → extract df['month']
OUTPUT:
  - Preprocessed DataFrame with sentiment and temporal features
4. EXPLORATORY DATA ANALYSIS (EDA)
  Plot:
    - Rating distribution (bar chart)
    - Average polarity per rating
    - Monthly review counts
    - Review length distribution (histogram)
    - Correlation heatmap (polarity, review length, rating)
OUTPUT:
  - Visual insights into sentiment, time patterns, and correlations
5. DEFINE FUNCTION: HBA_REC(df)
  INPUT:
    - Preprocessed DataFrame df
  PROCESS:
    - Clean text: lowercase, remove URL, non-alpha, stopwords
    - Recalculate polarity, subjectivity, review_len
    - Extract day, month, weekday, review_age from review date
    - Encode text using SentenceTransformer → apply PCA → get 'embedded'
    - Target Encode: brand, manufacturer, categories, colors
    - Normalize numerical features using MinMaxScaler
    - Combine all features → X
    - Apply BorderlineSMOTE to balance classes → X_bal, y_bal
  OUTPUT:
    - X: Full unbalanced feature set
    - y: Target labels
    - X_bal: Balanced features after SMOTE
```

- y_bal: Balanced labels
6. DEFINE FUNCTION: CAFRC(X, y)
  INPUT:
    - X: Unbalanced features
    - y: Labels
  PROCESS:
    - Apply RobustScaler to X → X_scaled
    - Train L1-regularized Neural Net → get l1_score
    - Train Autoencoder → get ae_score (reconstruction error)
    - Use GradientTape to compute Integrated Gradients → ig_score
    - Combine all three scores → combined_score
    - Rank features and select top 90% → selected_indices
    - Compute overall_feature_accuracy = (sum of selected scores / total scores) * 100
    - Select top features → X_selected
    - Apply SMOTE → get balanced X_selected and y_final
    - Train RandomForest with GridSearchCV on X_selected and y_final
  OUTPUT:
    - X_selected: Selected and balanced feature matrix
    - y_final: Balanced labels
    - selected_indices: Indices of top 90% important features
    - overall_feature_accuracy: % of total importance retained
    - combined_score: Feature importance vector
7. RUN FINAL PIPELINE
  - Call: X, y, X_bal, y_bal = HBA_REC(df)
  - Call: X_selected, y_final, selected_indices, overall_feature_accuracy, combined_score = CAFRC(X, y)
  Print:
    - Total number of selected features
    - Overall feature accuracy
  Plot:
    - Bar chart showing combined_score of selected features
OUTPUT:
  - High-quality, semantically rich feature set (X_selected)
  - Balanced and optimized label set (y_final)
  - Feature importance visualization
  - Trained RandomForest model ready for prediction
END

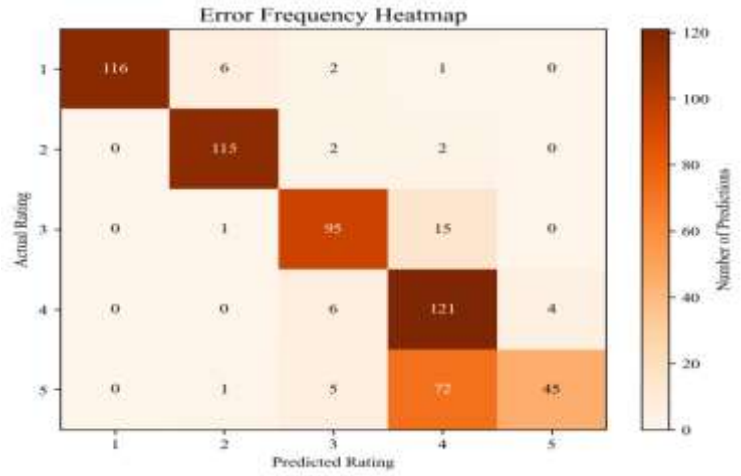
This workflow in algorithm 1 combines HBA-REC for behavior-aware review preprocessing and CAFRC for adaptive feature selection on Amazon product reviews. It cleans and encodes customer reviews and adds temporal and sentiment features. The data is then balanced using SMOTE. CAFRC ranks features by using L1 regularization, autoencoder error and integrated gradients. It selects the top features for classification. The final output is a compact and useful feature set that trains an optimized Random Forest model for later tasks.

IV. Results and Discussion

This section offers the comparative evaluation of preprocessing and feature selection techniques with common regression and classification performance metrics. The proposed new HBA-REC and CAFRC

approaches are compared with traditional and deep learning techniques. Experimental evidence indicates considerable performance gains, which establish the effectiveness of behavior-aware preprocessing and advanced feature selection in predictive accuracy improvement and robustness of models in e-commerce sentiment data.

Figure 4: Error Frequency Heatmap Chart



This figure 4 exhibits the distribution of predicted versus actual product ratings. Darker cells along the diagonal indicate many correct predictions; specifically, ratings 1, 2 and 4 agree at the actual star rating predicted star rating. The off-diagonal cells shows misclassification; for example, some true 5-star reviews are predicted as 4-star, indicating the model performs better with higher-than-expected ratings but appears to under-predict higher ratings as well. Overall, the density along the diagonal indicates overall good prediction performance with little dispersion of errors.

Figure 5: Distribution of Review Ratings Chart

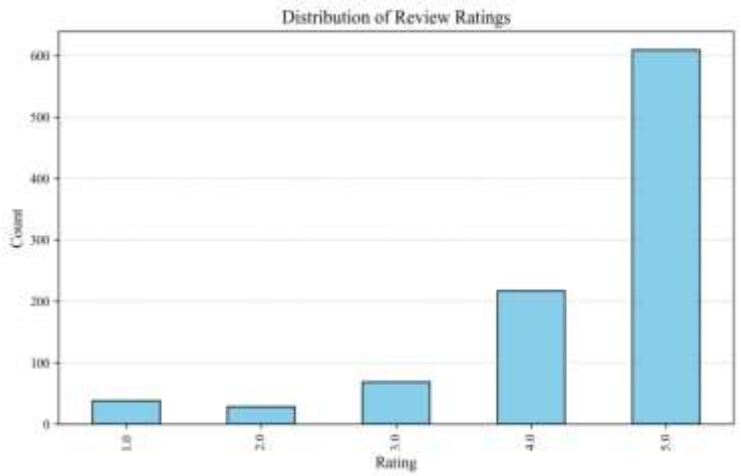


Figure 5 shows the distribution of review ratings in the dataset: It appears that the dominant review rating across all the users is 5.0, indicating the positive feedback overall. There is also a sizeable amount of ratings at 4.0, while very few users provided ratings in the lower end of the scale (1.0 to 3.0). This distribution indicates that there is a strong positive skew of customer reviews.

Figure 6: Monthly Review Counts Chart

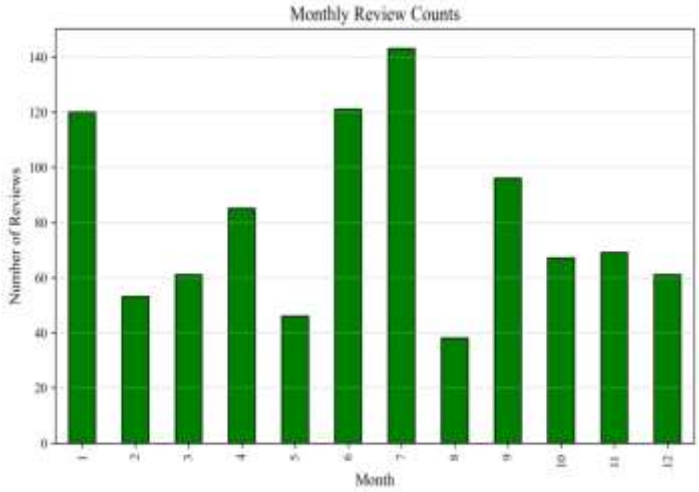


Figure 6 shows the quantity of product reviews submitted each month. The greatest numbers of reviews occurred in July month, June and January, of which these values seem to peak at the certain points of year suggest seasonal spikes. By contrast, August had the fewest review submissions. So the peaks in activity suggested that consumer engagement peaks mid-year and the beginning of calendar year.

Figure 7: Distribution of Review Lengths Chart

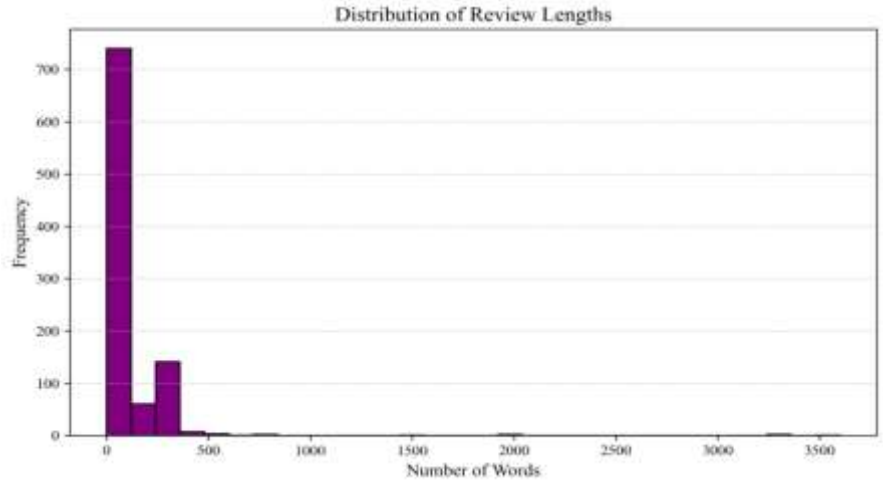


Figure 7 represents the distribution of review by the count of words. The bulk of reviews are short and mostly under 100 words with the highest frequency of review lengths in this range. There are very few

reviews of more than 500 words and extremely long reviews appear to be rare. These data suggest that the customers tend to prefer writing short feedback.

Figure 8: Correlation Heatmap Chart

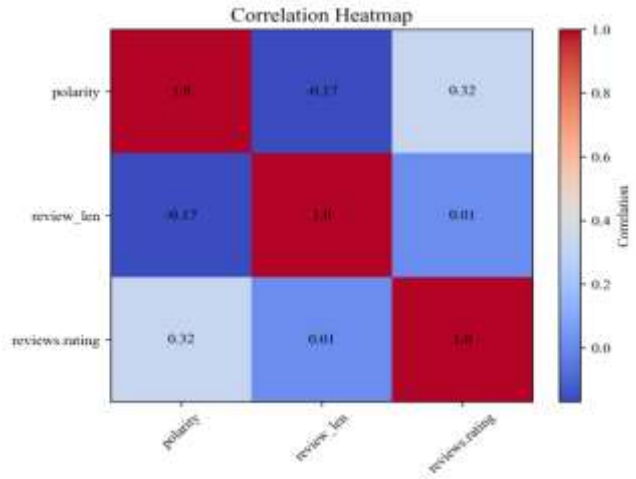


Figure 8 depicts the relationships between review polarity, length and rating. Even though polarity and rating have a moderate positive correlation (0.32), this indicates that a more positive sentiment is generally associated with higher rating. Review length correlates weakly negatively to polarity (−0.17) and nearly not-at-all with rating (0.01). Both of these correlation means that the length of review has no notable effect on the sentiment or rating.

Table 2: Preprocessing Performance Comparison Table

Methods	**MAE**	**MSE**	**RMSE**	**R^2**
Term Frequency-Inverse Document Frequency (TF-IDF) with SVM [34]	0.41	0.39	0.62	0.71
(ERF-XGB) [35]	0.34	0.30	0.55	0.79
LSTM-based Sentiment Regression [36]	0.31	0.26	0.51	0.83
HBA-REC (Proposed)	0.26	0.22	0.47	0.88

Table 2 compares the preprocessing methods based on regression performance metrics. The proposed HBA-REC method outperforms all the baseline approaches with the lowest Mean Absolute Error (MAE) (0.26), MSE (0.22) and RMSE (0.47). This shows that it provides more accurate and consistent sentiment predictions. It also achieves the highest R^2 score (0.88), which means it explains the most variance in target variable. Compared to the traditional methods like TF-IDF with SVM and deep models like Long Short Term Memory (LSTM), HBA-REC shows better behavior-aware preprocessing. It captures session dynamics and preserves rare interactions, leading to better model reliability and performance.

Figure 9: MAE Comparison Chart

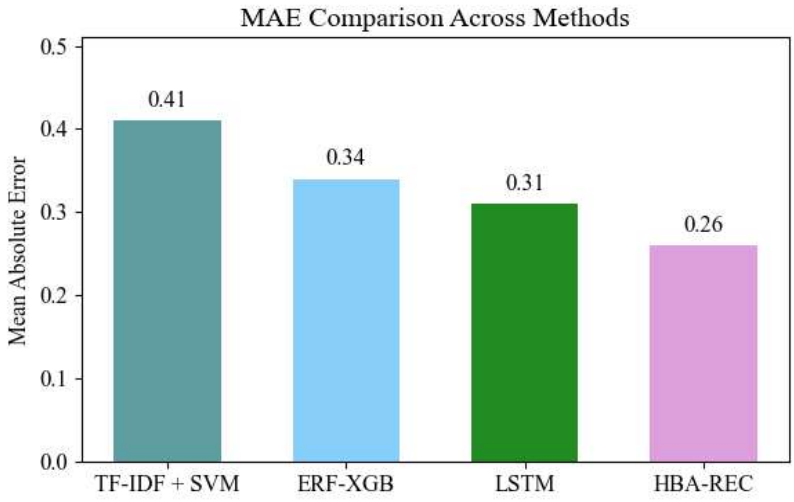


Figure 9 compares the MAE of four predictive methods. TF-IDF with SVM has the highest MAE at 0.41, which shows lower prediction accuracy. Extremely Randomized Forest-Extreme Gradient Boosting (ERF-XGB) and LSTM perform better with MAEs of 0.34 and 0.31, respectively. The HBA-REC method reaches the lowest MAE of 0.26, which indicates better performance and higher accuracy in predictions.

Figure 10: MSE Comparison Chart

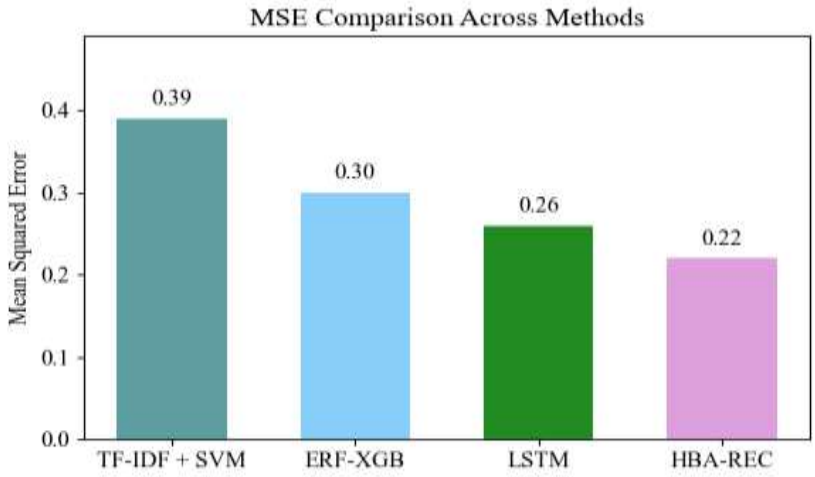


Figure 10 compares the Mean Squared Error (MSE) for four methods. TF-IDF with SVM has the highest error at 0.39, which shows the least accurate performance. ERF-XGB and LSTM have better results with MSE values of 0.30 and 0.26, respectively. The HBA-REC method achieves the lowest MSE of 0.22, which indicates its better prediction accuracy and reliability among the models tested.

Figure 11: RMSE Comparison Chart

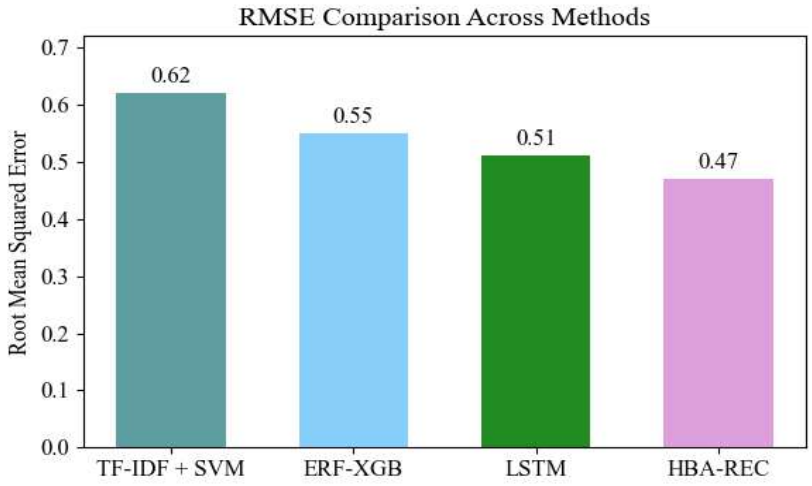


Figure 11 compares predictive ability of four approaches. TF-IDF with SVM has the maximum Root Mean Squared Error (0.62), meaning higher prediction variance. ERF-XGB and LSTM perform better with RMSEs 0.55 and 0.51, respectively. HBA-REC once again performs the best among all with minimum RMSE of 0.47, proving its higher prediction capacity and stability of the model.

Figure 112: R^2 Comparison Chart

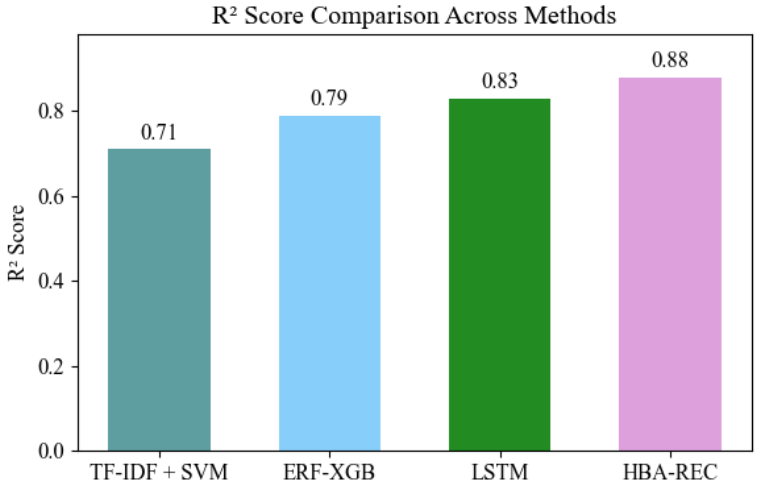


Figure 12 demonstrates the predictive power of each methodology. TF-IDF with SVM yields the lowest R^2 value of 0.71, suggesting that the model fit is relatively weak. ERF-XGB had an R^2 value of 0.79 and LSTM returned a 0.83 R^2, presenting performance in-between TF-IDF + SVM and HBA-REC. HBA-

REC achieved the highest R^2 score of 0.88, demonstrating the best accuracy and greatest correlation between the values predicted and the actual values.

Table 3: Feature Selection Method Comparison Table

Methods	Accuracy %	Total Features	Selected Features
Chi-Square [37]	86.72	27	18
PCA with Random Forest [38]	89.34	27	20
L1-Regularized Logistic Regression [39]	90.11	27	21
CAFRC (Proposed)	92.89	27	23

Table 3 compares the different feature selection methods based on classification accuracy and feature reduction effectiveness. The CAFRC method achieves the highest accuracy of 92.89%. It outperforms the traditional techniques such as Chi-Square (86.72%), PCA with Random Forest (89.34%), and L1-Regularized Logistic Regression (90.11%). All methods started with 27 features, but CAFRC retained only the most relevant 23 by using L1 regularization, autoencoder error and integrated gradients. This selection process is both clear and effective, improving model accuracy and interpretability. This makes CAFRC the best choice for handling complex e-commerce datasets.

Figure 13: Accuracy Comparison Chart

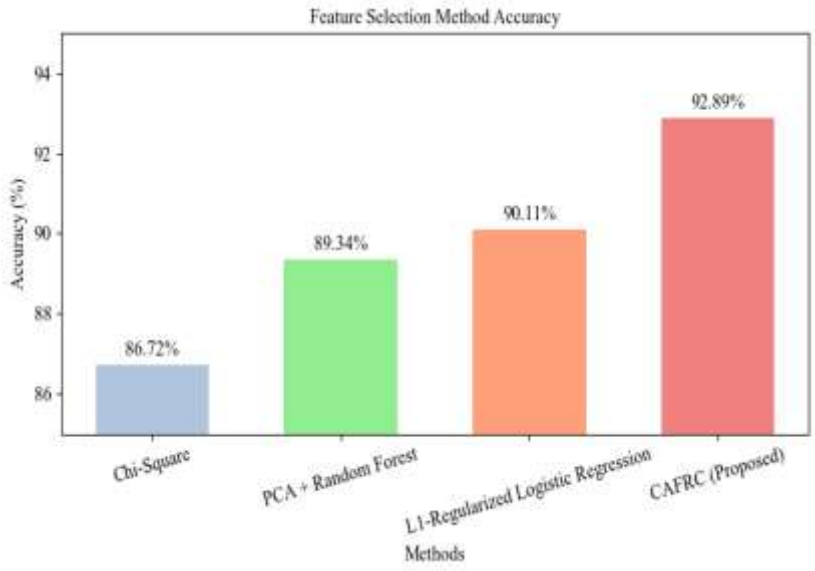


Figure 13 reflects the correctness of various feature selection techniques. The lowest correctness of 86.72% is registered by Chi-Square method, followed by PCA with Random Forest (89.34%) and L1-Regularized Logistic Regression (90.11%), reflecting modest improvements. The new method, CAFRC continues with the best accuracy of 92.89% as a higher performance in selecting the most meaningful features in order to have better models.

V. Conclusion

To sum up, the paper proposes an inter-related system that greatly improves the Amazon product review processing and analysis on the basis of combined use of Hierarchical Intra-Session Behavior Adaptive Reconstructions (HBA-REC) and CAFRC. This is because the HBA-REC technique sustains subtle user behavior through micro and meta sessions and enabling temporal and situational review dynamics to dominate over preprocessing. It preserves the amazing behavioral patterns and opinion changes, therefore the behavior of dataset. CAFRC augments this preprocessed information, then with feature space classical reconstruction about various layers as well as the choice of high utility features based on the mixture of autoencoder reconstruction loss and L1 regularization as well as integrated gradients. Results of the experiment prove that the HBA-REC framework is better than the classical preprocessing and deep learning-powered preprocessing on its predictive accuracy due to the lowest MAE, MSE and RMSE as well as the highest R2 score. Likewise, CAFRC also reaches a better feature selection accuracy in comparison to conventional approaches and demonstrates great potential in improving model robustness rather than overfitting. The proposed workflow as well provides a capability of interpretability, which provides transparency of the relevance of each level of the reconstruction feature. This qualifies it to be quite appropriate in sentiment classification, detection of fake reviews and product recommendation. When the HBA-REC and CAFRC are combined, these lead to effective behavior-aware pipeline having the high preservation of complex dynamics of the user reviews. This dual-framework approach increases the data quality and feature richness to provide a scalable platform toward further uses of advanced features in the ML applications within the e-commerce analytics environment. These models trained using this pipeline not only have a better accuracy but also have a better behavior across review manners. Providing real-time adaptation of HBA-REC and CAFRC in real-time review streaming scenarios has been investigated as a subject of future work to construct even more responsive e-commerce systems.

References:

1. Sharma, C., Kaur, A., Datta, P., & Gulzar, Y. (2025). Optimizing eCommerce Data: Effective Approaches for Data Collection, Cleansing, and Preprocessing. In Strategic Innovations of AI and ML for E-Commerce Data Security (pp. 1-30). IGI Global.
2. Chaudhary, N., & Roy Chowdhury, D. (2019). Data preprocessing for evaluation of recommendation models in E-commerce. Data, 4(1), 23.
3. Ahsain, S., El-Yusufi, Y., & Ait-Kbir, M. H. (2023). Optimizing Customer Experience Analysis Across Dataset Size Reduction and Relevant Features Selection. International Journal of Engineering Trends and Technology, 71(12), 78-89.
4. Peya, Z. J., Islam, M. S., Urme, N. J., & Auny, S. I. (2025, February). Machine Learning Based Potential Customer Recommendation System for E-Commerce Using Feature Selection and XAI. In 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.
5. Matuszelański, K., & Kopczewska, K. (2022). Customer churn in retail e-commerce business: Spatial and machine learning approach. Journal of Theoretical and Applied Electronic Commerce Research, 17(1), 165-198.
6. Wong, A. N., & Marikannan, B. P. (2020, December). Optimising e-commerce customer satisfaction with machine learning. In Journal of physics: Conference series (Vol. 1712, No. 1, p. 012044). IOP Publishing.
7. Ike, C. C., Ige, A. B., Oladosu, S. A., Adepoju, P. A., Amoo, O. O., & Afolabi, A. I. (2023). Advancing machine learning frameworks for customer retention and propensity modeling in ecommerce platforms. GSC Adv Res Rev, 14(2), 17.

8. Wijaya, D. R., Ibadurrohman, R. I. F., Hernawati, E., & Wikusna, W. (2024). Poverty prediction using E-commerce dataset and filter-based feature selection approach. Scientific Reports, 14(1), 3088.
9. Selvasundaram, K., Trivedi, P., Kasireddy, L. C., & Bhanawat, H. (2025). Artificial Intelligence in E-commerce and Banking: Enhancing Customer Experience and Fraud Prevention. Artificial Intelligence, 5(1).
10. Sina Mirabdolbaghi, S. M., & Amiri, B. (2022). Model optimization analysis of customer churn prediction using machine learning algorithms with focus on feature reductions. Discrete Dynamics in Nature and Society, 2022(1), 5134356.
11. Deniz, E., & Bülbül, S. Ç. (2024). Predicting customer purchase behavior using machine learning models. Information Technology in Economics and Business, 1(1), 1-6.
12. Tran, D. T., & Huh, J. H. (2023). New machine learning model based on the time factor for e-commerce recommendation systems. The Journal of Supercomputing, 79(6), 6756-6801.
13. Sarisa, M., Boddapati, V. N., Patra, G. K., Kuraku, C., Konkimalla, S., & Rajaram, S. K. (2020). An Effective Predicting E-Commerce Sales & Management System Based on Machine Learning Methods. Journal of Artificial Intelligence and Big Data, 1(1), 75-85.
14. Alghanam, O. A., Al-Khatib, S. N., & Hiari, M. O. (2022). Data mining model for predicting customer purchase behavior in e-commerce context. International journal of advanced computer science and applications, 13(2).
15. Demircan, M., Seller, A., Abut, F., & Akay, M. F. (2021). Developing Turkish sentiment analysis models using machine learning and e-commerce data. International Journal of Cognitive Computing in Engineering, 2, 202-207.
16. Mirdan, A. S., Baker, M. R., & Buyrukoğlu, S. (2025). Evaluating Machine Learning Performance and Consumer Sentiments on E-Commerce Platforms: A Comprehensive Twitter Analysis of Amazon. Ingénierie des Systèmes d'Information, 30(2).
17. Prabhakaran, N., & Nedunchelian, R. (2023). Oppositional Cat Swarm Optimization-Based Feature Selection Approach for Credit Card Fraud Detection. Computational Intelligence and Neuroscience, 2023(1), 2693022.
18. Farsi, S., & Chowdhury, M. (2025). EcomFraudEX: An Explainable Machine Learning Framework for Victim-Centric and Dual-Sided Fraud Incident Classification in E-Commerce. EAI Endorsed Transactions on Scalable Information Systems, 12(2).
19. Pustokhina, I. V., Pustokhin, D. A., Jayasankar, T., Jeyalakshmi, C., Díaz, V. G., & Shankar, K. (2021). Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. Information Processing & Management, 58(6), 102706.
20. Gupta, K., Jiwani, N., & Afreen, N. (2023). A combined approach of sentimental analysis using machine learning techniques. Revue d'Intelligence Artificielle, 37(1), 1.
21. Savci, P., & Das, B. (2023). Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages. Journal of King Saud University-Computer and Information Sciences, 35(3), 227-237.
22. Gupta, R. K., Hassan, A., Majhi, S. K., Parveen, N., Zamani, A. T., Anitha, R., ... & Muduli, D. (2025). Enhanced framework for credit card fraud detection using robust feature selection and a stacking ensemble model approach. Results in Engineering, 105084.
23. Daoud, Y., & Kammoun, A. (2024). Analyzing and forecasting e-commerce adoption drivers among SMEs: A machine learning approach. Human Behavior and Emerging Technologies, 2024(1), 7747136.
24. Vijayaragavan, P., Suresh, C., Maheshwari, A., Vijayalakshmi, K., Narayanamoorthi, R., Gono, M., & Novak, T. (2024). Sustainable sentiment analysis on E-commerce platforms using a weighted parallel hybrid deep learning approach for smart cities applications. Scientific Reports, 14(1), 26508.

25. Bagwari, A., Sinha, A., Singh, N. K., Garg, N., & Kanti, J. (2022). Cbir-dss: business decision oriented content-based recommendation model for e-commerce. Information, 13(10), 479.
26. Esmeli, R., & Gokce, A. (2025). An Analysis of Consumer Purchase Behavior Following Cart Addition in E-Commerce Utilizing Explainable Artificial Intelligence. Journal of Theoretical and Applied Electronic Commerce Research, 20(1), 28.
27. Jomthanachai, S., Wong, W. P., & Khaw, K. W. (2022). An application of machine learning regression to feature selection: a study of logistics performance and economic attribute. Neural Computing and Applications, 34(18), 15781-15805.
28. Alotaibi, F. M. (2023). A machine-learning-inspired opinion extraction mechanism for classifying customer reviews on social media. Applied Sciences, 13(12), 7266.
29. Alotaibi, S. R., Alkahtani, H. K., Aljebreen, M., Alshuhail, A., Saeed, M. K., Ebad, S. A., ... & Alotaibi, M. (2025). Explainable artificial intelligence in web phishing classification on secure IoT with cloud-based cyber-physical systems. Alexandria Engineering Journal, 110, 490-505.
30. Subramanian, R. S., & Prabha, D. (2022). Ensemble variable selection for Naive Bayes to improve customer behaviour analysis. Computer Systems Science & Engineering, 41(1), 339-55.
31. Sachin, D. (2015). Dimensionality reduction and classification through PCA and LDA. International journal of computer Applications, 122(17).
32. Deepa, B., & Ramesh, K. (2022). Epileptic seizure detection using deep learning through min max scaler normalization. International journal of health sciences, (I), 10981-10996.
33. Sağlam, F., & Cengiz, M. A. (2022). A novel SMOTE-based resampling technique trough noise detection and the boosting procedure. Expert Systems with Applications, 200, 117023.
34. Pakpahan, D., Siallagan, V., & Siregar, S. (2023). Classification Of E-Commerce Product Descriptions With The Tf-Idf And Svm Methods. Sinkron: jurnal dan penelitian teknik informatika, 7(4), 2130-2137.
35. Alghazzawi, D. M., Alquraishee, A. G. A., Badri, S. K., & Hasan, S. H. (2023). ERF-XGB: Ensemble random forest-based XG boost for accurate prediction and classification of e-commerce product review. Sustainability, 15(9), 7076.
36. Sharma, P., & Sagvekar, V. R. (2023). Weighted Ensemble LSTM Model with Word Embedding Attention for E-Commerce Product Recommendation. Journal of Communications Software and Systems, 19(4), 299-307.
37. Jiajie, G. (2024). Research on Mobile E-commerce Recommendation Algorithms Based on Logistic Regression Improved Model Features. Academic Journal of Engineering and Technology Science, 7(5), 111-115.
38. Jomthanachai, S., Wong, W. P., & Khaw, K. W. (2022). An application of machine learning regression to feature selection: a study of logistics performance and economic attribute. Neural Computing and Applications, 34(18), 15781-15805.
39. Mišić, V. V., Rajaram, K., & Gabel, E. (2021). A simulation-based evaluation of machine learning models for clinical decision support: application and analysis using hospital readmission. NPJ Digital Medicine, 4(1), 98.