Journal of Computing and Security

Research Article, 2026, Volume 13, Number 1 (pp. 35–57)

http://www.jcomsec.org

Behavioral Loyalty Prediction from Customer Reviews in E-Commerce: A Deep Learning and Explainable AI Approach

Pedram Kazemi Esfe [a] Fakhroddin Noorbehbahani [a,*]

[a] *Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran.*

ARTICLE INFO.

Article history:
Received: 03 December 2025
Revised: 01 April 2026
Accepted: 31 May 2026
Published Online: 21 June 2026

Keywords:
Loyalty Prediction, Customer Review Analysis, Natural Language Processing, RFM Model, and Explainable AI.

ABSTRACT

Understanding and predicting customer loyalty has become essential in increasingly competitive digital markets, yet conventional behavioral models, such as RFM, remain limited by their reliance on retrospective transactional data. Meanwhile, vast volumes of rich, interactive customer feedback, particularly online reviews, remain chronically underutilized for loyalty prediction. This study introduces a novel deep learning framework that predicts behavioral loyalty directly from customer review texts, eliminating the need for historical purchase records. Using a large e-commerce dataset of 58,841 samples, we combine advanced transformer-based language models (BERT, RoBERTa, ALBERT, GPT, and Qwen) with structured features and employ a genetic algorithm to optimize across 19,440 configurations of loyalty labeling schemes, neural architectures, text embeddings, and feature combinations. The resulting model achieves an accuracy of 0.92 and an F1-score of 0.90, substantially outperforming baseline machine learning methods and non-transformer text embedding approaches. Beyond performance, the integration of SHAP-based explainability provides transparent insights into the linguistic and behavioral cues driving loyalty predictions. These findings demonstrate that customer reviews contain powerful behavioral signals that can be transformed into actionable intelligence for real-time CRM, customer retention, competitive analysis, and multi-channel personalization, particularly in contexts where RFM data are incomplete or unavailable. This work offers both a methodological advancement in loyalty prediction and a practical, scalable tool for organizations seeking to leverage interactive feedback to build stronger, more proactive customer relationships.

1 Introduction

As the business environment becomes increasingly competitive and the number of substitutable goods and services rises, the study of customer loyalty has become more critical than ever. In such conditions, businesses must be able to examine and understand consumer behavior; it can be asserted that a business

* Corresponding author.
Email address: noorbehbahani@eng.ui.ac.ir(F. Noorbehbahani)
https://dx.doi.org/10.22108/jcs.2026.147672.1188
ISSN: 2322-4460

lacking satisfied and loyal customers will lack proper effectiveness [1, 2]. Furthermore, it is important to note that maintaining and improving a company's revenue status relies more on retaining current users than on acquiring new customers, making the analysis of current customer behavior and the prediction of their loyalty extremely valuable. The insights gained from this analysis can significantly aid business owners [3]. Business managers can leverage the results of customer loyalty analysis to identify the main reasons for customer dissatisfaction and take steps to address these issues, thereby enhancing customer satisfaction. Additionally, managers can use the findings to identify customers who are likely to discontinue using the company's products/services and allocate available resources more effectively to regain their satisfaction [4]. For example, offering promotions to loyal customers provides significant benefits, as such promotions directly enhance customer satisfaction, which in turn strengthens loyalty and encourages repeat business [5]. Loyal customers are more likely to respond positively to exclusive offers, such as gifts or VIP services, which not only reward their continued patronage but also differentiate the brand in a competitive market [5]. This cumulative effect of increased satisfaction and loyalty contributes to sustained business growth and long-term profitability.

Customer loyalty is commonly divided into behavioral and attitudinal forms. Behavioral loyalty refers to observable actions, such as repeated purchases or frequent interactions with a brand, often measured through buying frequency, spending level, and customer tenure. It may also be influenced by switching costs, where customers remain with a brand due to the perceived inconvenience or expense of changing providers rather than true satisfaction [6]. Attitudinal loyalty reflects the emotional and psychological attachment customers feel toward a brand, captured through their attitudes, preferences, and willingness to recommend it. Factors such as satisfaction, brand image, and relationship quality shape this form of loyalty. Although attitudinal loyalty can lead to behavioral loyalty, the reverse is not always true, as customers may repurchase without strong emotional commitment [6]. Both forms matter, but behavioral loyalty is often viewed as more critical due to its direct link with sales and long-term business sustainability.

The rise of online services has generated vast volumes of unstructured textual data. Natural Language Processing (NLP) techniques can convert this raw information into structured formats, improving human-computer interaction and enabling diverse analytical applications [7]. Despite the recognized importance of customer loyalty and the availability of such data in online retail contexts, few studies have applied textual analysis to predict loyalty. As a result, much of this unstructured data remains underutilized, even though NLP and machine learning methods could effectively estimate customer loyalty levels.

The RFM model is highly effective and valuable for predicting customers' future behavioral responses, assisting business owners in achieving greater profitability in a short period. This model can efficiently identify a company's valuable customers and provide critical insights to owners and managers. Moreover, it enables business owners to define timely promotions for the right customers by identifying patterns among them. However, the RFM model is retrospective and requires historical data that may not always be available. In some situations, the necessary information for constructing this model may be lacking. Therefore, developing a model that can predict customer loyalty based solely on available data, such as customer reviews, would be immensely beneficial and practical. Such a model could prove useful in the following scenarios:

- New Customers: For newly acquired customers whose reviews are available, but for whom sufficient data to calculate RFM metrics is lacking.
- Competitor Analysis: When customer reviews of a competitor are publicly accessible, but the information required to calculate RFM is not available, the model can analyze customer loyalty in this context.
- Social Media Contexts: In platforms like social media, where customer information may not be uniformly collected for RFM scoring, this model can assess customer loyalty solely based on their reviews.

Explainable AI (XAI) comprises methods that make complex machine learning models, particularly "black-box" systems, transparent and interpretable to humans. Beyond generating predictions, XAI explains how and why a model makes certain decisions through feature-importance scores or visualization of internal mechanisms [8, 9]. This interpretability is essential for trust, accountability, and actionable insights in business applications. In customer relationship management (CRM), XAI helps translate model outputs into strategies for acquisition, retention, and service optimization. In text-based analytics, methods such as SHAP and LIME identify which words or sentiments most influence satisfaction, churn risk, or loyalty. By offering local and global explanations, for instance, highlighting phrases like "late delivery" or "poor support" that reduce loyalty, XAI guides marketing and customer service teams toward targeted interventions.

Applying XAI to CRM text analysis thus enables

firms to uncover the main drivers of sentiment and loyalty. Insights such as "fast shipping" elevating loyalty or "faulty packaging" predicting churn allow businesses to tailor actions for specific segments, retrain staff to address frequent pain points, and personalize offers based on satisfaction factors. Combined with NLP, XAI transforms CRM models from merely predictive to truly prescriptive, helping organizations act on the most influential determinants of customer engagement.

This study aims to predict customer behavioral loyalty in an online store by integrating RFM-based segmentation with textual review analysis. Using an e-commerce dataset, we first categorize customers into loyalty classes via various RFM configurations and vectorize their reviews using multiple language models. Our objective is to identify the optimal combination of RFM labeling, language models, class counts, and machine learning algorithms for loyalty prediction. To ensure interpretability, we employ SHAP for local and global explanations. This approach identifies specific textual drivers, such as individual words or recurring sentiment patterns, that influence loyalty predictions, providing stakeholders with actionable data-driven insights.

The structure of the paper is organized as follows. Section 2 reviews previous studies in the field of customer loyalty prediction, highlighting the gaps that this study seeks to address. Section 3 explains the proposed method to predict customer loyalty through data processing and modeling of various features, alongside the application of natural language processing techniques. Section 4 presents the results and interpretations obtained after implementing the proposed methodology. Finally, Section 5 offers a comprehensive summary of the paper along with limitations and recommendations for future research.

2 Related Works

This section presents a comprehensive review of studies focused on predicting and analyzing customer loyalty. The studies are categorized based on the type of data utilized (structured or unstructured).

2.1 Research Utilizing Structured Data

Kishada et al. used an Artificial Neural Network (ANN) to predict customer loyalty in Malaysian Islamic banking, achieving a correlation coefficient of 0.98. Data was labeled via survey questionnaires, identifying satisfaction and service quality as key predictors [10]. Alizadeh and Bidgoli proposed a hybrid data mining model, labeling customers via K-means clustering based on spending levels. Applying various classifiers, their Bagging-ANN approach achieved the highest prediction accuracy of 0.71 [11]. Machado et al. compared Gradient Boosting models for a South American financial company, finding that LightGBM outperformed XGBoost in RMSE efficiency. Loyalty scores derived from transaction histories served as the target labels [12].

Wijaya et al. evaluated data mining algorithms for telecom churn prediction, with customer labels derived from behavioral indicators like usage frequency. The C4.5 algorithm yielded the highest accuracy at 0.81 [13]. Sulistiani et al. utilized a Dynamic Mutual Information and Support Vector Machine (DMI-SVM) model on survey-labeled data; the nonlinear model incorporating satisfaction outperformed Naïve Bayes [14]. Figueroa applied self-organizing maps (SOM) for RFM segmentation, followed by expert labeling and a multilayer perceptron (MLP) classification in Chilean retail, identifying 19% of the sample as loyal [15]. Granov utilized clustering and Logistic Regression for a fashion e-commerce company, classifying transaction-labeled customers with accuracies of 0.68 for churners, 0.75 for returners, and 0.98 for loyalists [16]. Lee et al. combined K-means clustering of RFM attributes with supervised classification. Labeling the "top 10% spenders" as the target, their Decision Tree model achieved 0.77 accuracy [17]. Abbassy assessed e-commerce loyalty predictors, highlighting Gradient Boosting's high accuracy of 0.88 on data categorized strictly as retained or churned [18]. Qin predicted e-commerce customer churn across three loyalty levels using an optimized XGBoost algorithm, effectively handling imbalanced data with an F1-score of 0.90 [19]. Sulistiani and Tjahyanto categorized instant noodle questionnaire respondents as loyal or disloyal, achieving 0.83 accuracy using Random Forest with Chi-Square feature selection [20].

Wijaya et al. predicted telecommunications churn by categorizing reasons for unsubscription. Utilizing 2,269 records, the C4.5 algorithm achieved a 0.81 accuracy rate [21]. Syukur et al. integrated bootstrapping, Weighted Information Gain, and SVM grid search for telecom loyalty prediction. Using pre-labeled service duration datasets, they achieved accuracies of 0.93 and 0.99 [22]. Tong et al. clustered telecom customers using k-means and labeled them via Net Promoter Scores (NPS). Their XGBoost decision tree model yielded an accuracy of 0.82 (training) and 0.71 (testing) [23]. Wassouf et al. used the TFM model to label loyalty. Their binary classification model achieved an accuracy of 0.87 and AUC of 0.80, while multi-class predictions yielded 0.73 precision and 0.74 recall [4].

Hamdan and Othman analyzed hotel loyalty via

predefined dataset attributes, finding the decision tree algorithm yielded the highest accuracy (0.71) and identifying "discount" as the prime predictor [24]. Hsu et al. modeled resort tourism loyalty using a LISREL-integrated Bayesian Network. Validated by survey data, it outperformed traditional neural networks with lower mean average errors [25]. Hamdan et al. utilized hotel reservation features to label loyalty via purchasing behaviors, attaining a model accuracy of 0.71 and an F1-score of 0.74 [26]. Singh predicted airline loyalty using questionnaire-labeled passenger data; utilizing an Artificial Neural Network, the model reached 0.89 accuracy and identified satisfaction as the key mediator [27].

Wu et al. predicted "class loyalty" and plagiarism in an online programming course using convolutional neural networks (CNNs) on student behavior logs, reaching 0.95 accuracy for class loyalty predictions [28]. Wang et al. proposed an updated Recency, Frequency, and Duration (RFD) model to classify mobile app users into four loyalty segments, discovering that most users maintain their initial loyalty status over time [29].

2.2 Research Utilizing Unstructured Data

Noorbehabani et al. predicted attitudinal customer loyalty for a Persian online store by analyzing product reviews, labeling data based on customers' willingness to recommend. They achieved an initial accuracy of 0.67, which improved to 0.89 in a binary classification setting [30]. Ghani et al. measured loyalty using aggregated sentiment scores from a large dataset of Amazon reviews. Employing NLP preprocessing and a fuzzy logic model, they achieved 0.94 accuracy, though the study conflates sentiment polarity directly with customer loyalty [31].

Shahid et al. integrated sentiment analysis and user experience data from 17,000 respondents to predict airline loyalty. Their XGBoost model achieved an 85% success rate in predicting return visits. However, the study's labeling method (based solely on continued service use) and its exclusive focus on emotion-based text analysis were noted as overly simplistic [32]. Khan and Urolagin measured attitudinal loyalty by categorizing customers based strictly on explicit declarations of loyalty or disloyalty in tweets. This approach, which relied solely on sentiment polarity, may be imprecise as it disregards customers who do not explicitly state their stance [33].

Kilimci designed a model to predict mobile app user satisfaction, achieving an accuracy of 0.86. While the study uses satisfaction to estimate loyalty, noting that highly loyal users provided more reviews and positive ratings, it acknowledges that satisfaction and loyalty, despite a strong correlation, are not identical [3]. Urolagin and Patel introduced new social data metrics (such as the User Sentiment Score) to predict loyalty for online learning platforms using Twitter data. By classifying customers solely on tweet sentiment polarity, their Dense Neural Network (DNN) achieved a 0.98 accuracy [34]. Zaki et al. modeled behavioral loyalty for a UK heavy equipment company using surveys, purchase data, sentiment scores, and RFM/NPS metrics. Without setting a strict boundary between loyal and non-loyal customers, their predictive model achieved an accuracy of 0.98 and an AUC of 0.99 [35].

2.3 Findings from Related Works

There have been studies where researchers used direct responses from customers (such as through surveys or the NPS metric) to label customer loyalty and then utilized structured data to predict these labels [10–13]. Researchers who have incorporated unstructured data in their methods have also typically relied on attitudinal loyalty, which reflects a customer's willingness to recommend products [3, 14–16]. There have been studies that used RFM and its variants to analyze customer loyalty, but none of them utilized unstructured data [4, 17–19]. No research has successfully employed optimization and labeling to accurately predict behavioral loyalty from unstructured data.

Conversely, Zaki et al. aimed to assess behavioral loyalty through purchasing patterns. However, the study defines a customer as a "churner" if their recency (days since last purchase) exceeds the overall average inter-purchase interval (88.1 days). This single global cut-off treats all customers identically, ignoring differences in typical purchase rhythms by segment, product line, seasonality, or contract structure. In practice, this can mislabel slow but still-loyal customers as churners, and vice versa. Moreover, the "active customer" binary feature (loyal vs. churner based on the same 88-day rule) is used both as an input to the model and to construct the output label. This creates a tautological predictor that directly encodes the target, inflating apparent accuracy. Additionally, this study relied solely on sentiment polarity from review texts and did not further utilize this unstructured data, meaning that the textual data was only used to assess sentiment scores, with no other insights extracted from the content [35].

The study by Shahid et al. [32] derives loyalty labels from the same sentiment thresholds used during testing, rather than independent measures like repeat purchases or surveys. As a result, its reported 94% accuracy mainly reflects consistency with its own rules, not true predictive power. Because evaluation relies on labels generated by the same fuzzy-logic criteria,

it becomes circular. Moreover, focusing only on sentiment polarity ignores key aspects of real loyalty, such as repurchase behavior, referrals, and long-term commitment, leading to a limited and potentially misleading metric. A key limitation of Shahid et al. [32] is its narrow definition of loyalty, based solely on whether a customer repurchased within a year. While this captures basic repeat behavior, it overlooks the multi-dimensional nature of loyalty described in models like RFM, which consider recency, frequency, and spending. This simplification may miss customers with strong engagement who did not repurchase within the timeframe. Additionally, the study uses LIWC-based sentiment scoring without leveraging advanced NLP or large language models, limiting its ability to extract deeper insights from customer text such as intent, emotion, and nuanced behavioral signals.

In summary, existing studies on customer loyalty prediction generally fall into two main categories. The first group relies on structured transactional data, particularly the RFM model and its variants, to capture behavioral loyalty. While effective, these approaches are inherently retrospective and require detailed historical purchase data, limiting their applicability in scenarios such as new customers, competitors, or social media contexts. The second group utilizes unstructured textual data, often focusing on sentiment analysis to estimate attitudinal loyalty. However, these methods typically oversimplify loyalty by equating it with sentiment polarity and fail to capture actual behavioral patterns.

In contrast, our proposed approach bridges this gap by integrating both perspectives. Unlike RFM-only studies, it leverages rich textual information from customer reviews to enable loyalty prediction even when transactional data is unavailable. At the same time, unlike text-only approaches, it grounds the prediction in behavioral loyalty by generating labels using the RFM model rather than relying solely on sentiment. Furthermore, this study goes beyond prior work by incorporating advanced transformer-based language models to capture deep semantic patterns in text and by applying a genetic algorithm to systematically optimize model configurations. This combination allows for a more comprehensive, accurate, and generalizable framework for customer loyalty prediction, effectively overcoming the limitations of both traditional RFM-based and sentiment-based approaches. Moreover, to the best of our knowledge, no existing studies have employed explainable AI methods to interpret and visualize the predictions of black-box models specifically for customer loyalty prediction, making this study the first to bridge that gap in explainability.

3 Proposed Method

This study proposes a unified framework for predicting customer behavioral loyalty from review data by integrating structured transactional information with unstructured textual data. The overall objective is to develop a model that can accurately infer customer loyalty based on review content when traditional behavioral indicators (such as RFM metrics) are unavailable.

At a high level, the proposed method consists of four main components: (1) data preparation and feature extraction, (2) textual data processing and vectorization using advanced language models, (3) behavioral loyalty labeling using the RFM model, and (4) predictive modeling and optimization. These components are interconnected such that transactional data is first used to generate reliable loyalty labels, while review texts are transformed into numerical representations and combined with structured features to train predictive models. The final stage employs a genetic algorithm to identify the optimal combination of model configurations.

More specifically, the employed dataset plays a dual role in the framework. Transactional data (e.g., purchase history) is used to compute RFM metrics and generate behavioral loyalty labels, which serve as the ground truth for supervised learning. At the same time, customer reviews (text, titles, and scores) are processed and converted into feature representations that act as inputs to the predictive models. This design enables the model to learn the relationship between textual expressions and underlying behavioral loyalty. Figure 1 illustrates the proposed loyalty prediction method, comprising 11 stages. Stages 5 through 10 involve the iterative execution and testing of various parameters using an optimization algorithm. The steps of the proposed method are explained in more detail below.

3.1 Dataset Preparation

In this stage, the dataset was prepared, and order-related features were extracted to segment customers using the RFM model and assign loyalty labels. Additionally, the review texts, review titles, and review scores were extracted from the dataset. These features were subsequently used to develop various loyalty prediction models.

The dataset used in this study is the Brazilian E Commerce Public Dataset by Olist, which contains multiple related tables describing different aspects of an e commerce platform, including customers, products, sellers, orders, payments, and reviews. The dataset comprises 100,903 entries and was obtained

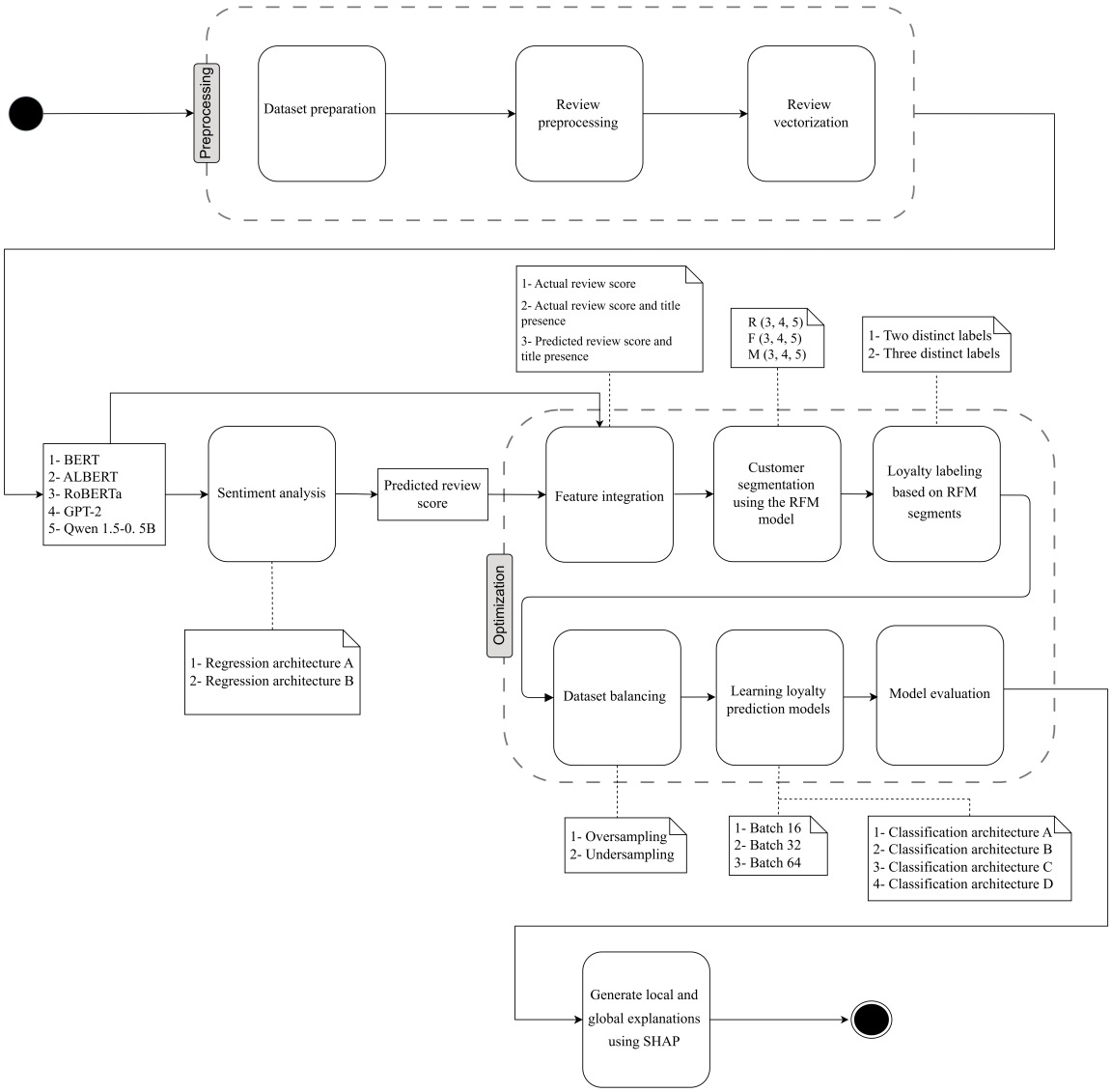


Figure 1. Proposed Method.

from the official Kaggle website [1] . The Olist customers dataset contains basic information about customers, such as customer IDs and their geographic location (city and state). Geographic coordinate information based on zip codes is provided in the Geolocation dataset, which can be used for spatial analysis of customers and sellers. The Olist Sellers dataset includes information about sellers, such as seller IDs and their locations. Product-related attributes, including product ID, category, weight, and dimensions, are stored in the Products dataset, while the product category name translation table provides English translations of product category names originally recorded in Portuguese.

Order-related information is distributed across several tables: Orders dataset contains general order information such as order ID, customer ID, purchase timestamp, order status, and delivery dates; Order items dataset records the individual items within each order, including the product, seller, price, and shipping cost; and Order payments dataset includes details about payment methods, payment installments, and the amount paid for each order. Finally, customer feedback is captured in the review tables. The Olist order reviews dataset contains review titles, review texts, and review scores provided by customers, while the Order reviews translate dataset provides translated versions of these reviews to facilitate text analysis in English.

After merging these tables, the features required for this study have been extracted, which are explained in Table 1.

[1] https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Table 1. Dataset Description.

Feature	Description	Usage
Review text	Customer reviews detailing their experience after completing an order.	Customer review text has been transformed into numerical vectors for building and training predictive models of customer behavior.
Review score	Customer ratings of their order experiences on a 1 to 5 scale.	These scores served as the foundation for developing models that predict customer behavioral loyalty. They were also used as labels for sentiment analysis models.
Order completion date	The completion date of customer orders is recorded in this field.	This field is used to calculate the Recency (R) in the RFM model. For each customer, the date of their most recent order is identified, after which these dates are ranked. Customers are then divided into equal-sized groups according to the Recency parameter set in the model.
Customer identifier of orders	Unique identifier for the customer associated with the order.	This field was used to calculate the Frequency (F) in the RFM model by counting the completed orders per customer, ranking customers by this count, and dividing them into equal-sized groups based on the Frequency parameter.
Order Values	Monetary value of completed orders.	The average order value for each customer is calculated and ranked to determine the Monetary (M) in the RFM model. Customers are then divided into equal-sized groups based on this ranking and the calculated Monetary parameter.
Title Presence	Indicates the presence or absence of a title in reviews.	This feature is used as an input feature to machine learning algorithms for predicting customer loyalty.

3.2 Review Preprocessing

In natural language processing tasks, text preprocessing is necessary. This process is used to prepare textual data before analysis and includes various steps, such as tokenization, stop word removal, and stemming.

After retrieving the text and titles of the reviews along with their numerical scores, the data was preprocessed to build models for predicting customer loyalty labels. Both the review scores and the predicted scores were normalized to a scale of 0 to 1 using min-max scaling based on the dataset's minimum and maximum values.

In the next step, for text preprocessing, all reviews are translated from Brazilian Portuguese to English using Google's machine translation system. This is done to facilitate better monitoring of the text analysis and processing, as well as to leverage stronger models available in English. To ensure the accuracy and reliability of this translation process, a study conducted by Avelino et al. [36] as reviewed, in which researchers used an innovative method to evaluate and compare the performance of various machine translation systems in translating from Brazilian Portuguese to English. The study concluded that Google's machine translation system had the best performance in this context, with an accuracy of 0.85, compared to other prominent machine translation systems. Next, all non-alphabetic characters are removed from the review text, and all uppercase letters in the words are converted to lowercase. Additionally, all English stop words present in the reviews are removed.

Finally, all words in the review texts have been reduced to their root forms. Since lemmatization produces more accurate and context-aware results compared to stemming, it is expected to perform better in tasks such as text classification. In this study, lemmatization has been used for root word extraction. For this purpose, words have been reduced to their root forms in the dictionary using their syntactic roles in sentences using WordNet [37].

3.3 Review Vectorization

The tokenization and vectorization process employed model-specific tokenizers from BERT [38], RoBERTa [39], GPT-2 [40], and Qwen 1.5-0. 5B [41]. These language models demonstrate advanced capabilities in capturing syntactic structures, semantic relationships between words, and contextual understanding of sentences. Qwen, developed by Alibaba Cloud, features a hybrid architecture that combines dense and Mixture-of-Experts (MoE) layers, enabling efficient scaling to trillion-parameter sizes while supporting multilingual

tokenization and dynamic context windows extending up to 32k tokens [41]. GPT, a purely autoregressive language model, utilizes unsupervised pre-training on large-scale text corpora to generate coherent text sequences, with its 1.5-billion-parameter architecture enabling robust performance across diverse linguistic tasks without task-specific fine-tuning [40].

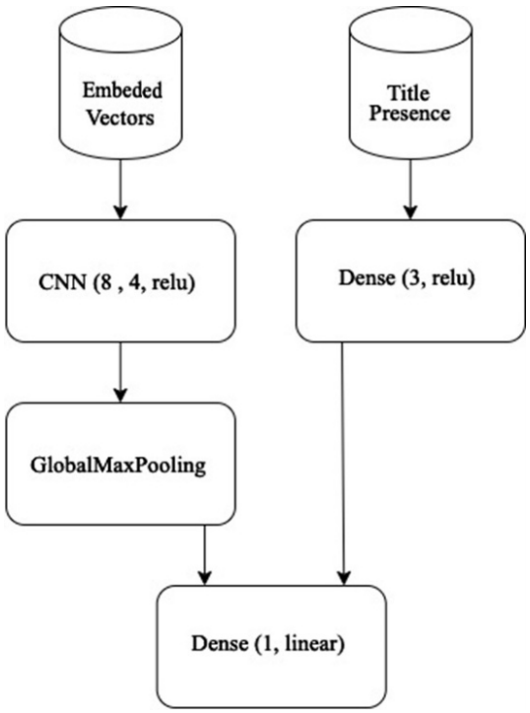
For the vectorization process, tokenized outputs were transformed into numerical embeddings using bidirectional attention mechanisms in BERT and RoBERTa, autoregressive sequential modeling in GPT, and Qwen's domain-adaptive tokenization that dynamically processes code-switched text while reducing dependency on labeled data through parameter-efficient fine-tuning techniques. GPT's architecture, based on transformer decoder layers, excels at modeling sequential dependencies in text through next-token prediction, making it particularly effective for generating syntactically coherent embeddings. Qwen enhances multilingual processing and long-context handling through its innovative MoE architecture, addressing challenges posed by code-mixed or lengthy customer reviews.

The use of Qwen and GPT offers significant advantages for text analysis applications. Qwen's architectural scalability and multilingual capabilities improve processing of code-mixed or lengthy review content, while GPT's unsupervised training on diverse text corpora (including books, articles, and web content) provides broad generalization for semantic feature extraction. Comparative evaluations demonstrate that both models surpass earlier architectures in few-shot learning scenarios, as evidenced by performance benchmarks in the Qwen Technical Report [41], and analyses of GPT's zero-shot adaptability in cross-domain tasks [40].

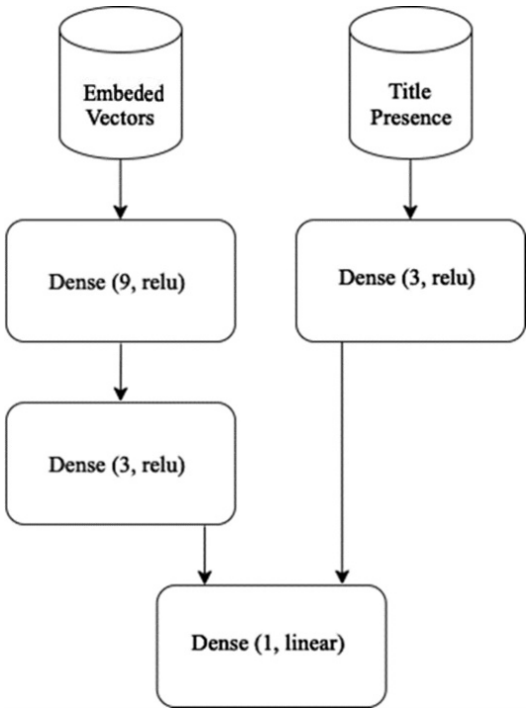
In this study, to utilize these language models for text vectorization, texts are first tokenized using a trained tokenizer with five different language models: BERT, ALBERT, RoBERTa, GPT, and Qwen. These tokens are then converted into numerical vectors using these language models. These numerical vectors can later be used to build and train predictive models for customer loyalty.

3.4 Sentiment Analysis

Models have been developed for sentiment analysis of customer reviews that can predict review scores using the text of the reviews and the presence of the review title. Since there may be situations where access to actual scores accompanying the reviews is not available, an optimization algorithm is used to build and evaluate the final model by relying on both predicted scores and actual ones. This approach allows the performance of the model to be evaluated under such conditions.



(a) Regression Architecture A.



(b) Regression Architecture B.

Figure 2. Two Architectures Applied for Sentiment Analysis.

Since these models predict numerical scores along with reviews, they have been implemented as regression models. For this purpose, ten different models were created by combining two different architectures

to build regression models with five language models: BERT, ALBERT, RoBERTa, GPT, and Qwen. Then, the R^2 metric was calculated for them using 5-fold cross-validation. Based on the obtained results, the best model was selected and used to calculate the sentiment scores. Since this is a regression problem, a simple linear activation function has been used in the final layer of both architectures. The two architectures employed are shown in Figure 2.

3.5 Feature Integration

At this stage, two structured features are stored alongside the vectors generated from review texts using language models in a new dataset. These structured features include:

- Title Presence: If a customer has never entered a title for any of the reviews related to their orders, this feature is set to 0. Otherwise, if they have entered a title for at least one of their reviews, this feature is set to 1.
- Actual Review Score: Upon completing an order and filling out the questionnaire, the customer is asked to provide a numerical score between 1 and 5.

In the optimization algorithm employed, the following three different scenarios were used for constructing predictive models based on these features:

(1) Using only the Actual Review Score.
(2) Using the Actual Review Score along with the Title Presence.
(3) Using the predicted score of the review along with Title Presence.

The reason for considering the Title Presence in constructing predictive models for reviews is based on the initial hypothesis of the study. It was assumed that if a customer is loyal to a company, they might be more careful and patient when writing a review, and thus, they might also add a title to their review. Therefore, there might be a relationship between the presence or absence of a title in customer reviews and their loyalty. To test this hypothesis, in one scenario, this feature was used to build loyalty prediction models, while in another scenario, it was not used. The results obtained from testing this hypothesis are presented in Section 4.

3.6 Customer Segmentation Using the RFM Model

Since this dataset includes customers who have placed multiple orders from this online store over time, it is necessary to aggregate the order information for each customer before labeling customer loyalty using the RFM model. In this study, weighted averaging was employed for this purpose. The numerical vectors associated with customer reviews, as well as the numerical scores accompanying those reviews (both actual and predicted), were averaged in such a way that more weight was given to more recent orders. This is because the most recent order and the customer's latest opinion about the store are more important and closer to their current view than earlier ones.

Next, different variations of the RFM model were used for customer segmentation. To achieve this, all possible combinations of predefined values for the three parameters R, F, and M were utilized to segment customers, and the results were saved in separate files. The segmentation process was carried out by first sorting and then dividing customers based on each of the three mentioned parameters separately. Then, by combining the resulting segments, the RFM loyalty score for each customer was calculated. The predefined values for each of the RFM model parameters range from 3 to 5.

3.7 Loyalty Labeling Based on RFM Segments

After segmenting customers using the RFM method and sorting the customers in descending order based on these segments, the customers were divided into different groups to create homogeneous loyalty groups and label their loyalty.

In this research, two main settings were considered for loyalty labeling:

- Two-class labeling: In this setting, the sorted list of customers was split exactly into two halves. Customers in the top half were classified as "Loyal", while those in the bottom half were classified as "Non-loyal".
- Three-class labeling: In this scenario, the sorted list of customers was divided into three equal parts. This more refined division allowed for customers to be categorized into three groups: "Loyal", "Neutral", and "Non-loyal".

3.8 Dataset Balancing

The distribution of data across loyalty classes was uneven, with some classes having significantly fewer samples compared to others. To balance the dataset and improve model performance, two common sampling techniques, Oversampling and Undersampling, were applied in this study. In the Oversampling method, random samples from the minority class (the class with the fewest data) are duplicated to match the number of samples in the majority class. In other words,

by creating random copies of the existing samples in the minority class, the number of samples in this class is increased. This method is particularly useful when the number of samples in the minority class is very small, and there is little concern about introducing noise into the data. In the Undersampling method, random samples from the majority class are removed to match the number of samples in the minority class. This method reduces the data volume and may lead to the loss of valuable information. In this study, the commonly used methods of Random Oversampling and Random Undersampling were applied.

3.9 Learning Loyalty Prediction Models

In the next stage, deep neural networks were developed using structured features (Title Presence and actual/predicted review scores) alongside review text vectors. To prevent the numerical features from being overshadowed by the larger text embeddings, separate branches were designed to process these features independently. The outputs of both branches were then combined, enabling the use of both feature types in the model. Four distinct feedforward neural network architectures were employed to model customer loyalty, as shown in Figure 3.

During training, the model is fed data for 200 epochs, but to prevent overfitting and save resources, training stops if the validation loss doesn't improve after 25 consecutive epochs. The ADAM optimizer is used, with Categorical Cross Entropy (CCE) for multi-class customer loyalty classification and Binary Cross Entropy (BCE) for binary classification. Convolutional Neural Networks (CNNs), commonly used for image recognition, are also applied to text classification and sentiment analysis in natural language processing. Therefore, CNN is also utilized in our study.

We applied TabTransformer in our study because it excels at processing tabular data by leveraging transformer architectures to create contextual embeddings for categorical features. Its self-attention mechanisms capture dependencies and interactions among categorical columns, allowing for richer feature representations. By combining these embeddings with numerical features, TabTransformer outperforms traditional methods in modeling complex feature relationships and making accurate predictions [42].

In architectures A and B, the first branch processes review text embeddings through a one-dimensional convolutional layer with a window size of 3 and 6 windows, followed by ReLU activation, batch normalization, and pooling. The second branch feeds numerical features into a dense layer with 3 nodes and ReLU activation. The outputs of both branches are combined and passed through a final dense layer with one node and a Sigmoid activation for binary classification, or three nodes with Softmax for a three-class problem, to predict customer loyalty.

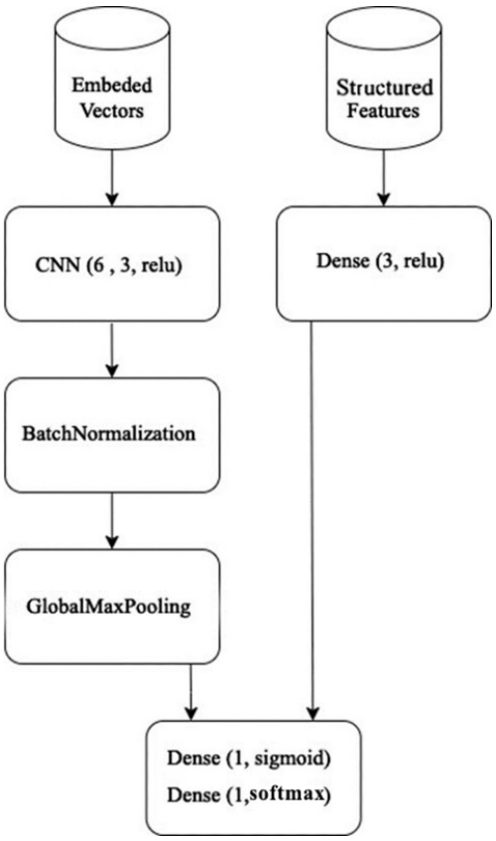
Transformer models like BERT, RoBERTa, ALBERT, Qwen, and GPT excel in customer loyalty prediction by handling both numerical and short-text data. Their self-attention mechanisms capture semantic patterns in brief texts (e.g., reviews or survey responses), and when combined with CNNs, Dense layers, or TabTransformer, they create hybrid models that optimize both text and numerical features. This multimodal approach scales well with large datasets, reduces training time via transfer learning, and avoids overfitting. It also mitigates the "curse of dimensionality" by generating task-specific embeddings for text data before combining them with numerical input.

As shown in Figure 3, Architecture A has two branches: one uses convolutional neural networks for review text embeddings, while the other processes structured features with dense layers. Architecture B increases the complexity of the convolutional layer by using a larger window size (64) and 7 windows. Architecture C adds two dense layers with ReLU activation to the embedding-processing branch. Architecture D incorporates a dense layer (32 nodes) in the text embedding branch and a TabTransformer module in the numerical feature branch, followed by a dense layer (16 nodes). All architectures have a final dense layer with a sigmoid activation for binary classification or Softmax for three-class scenarios.

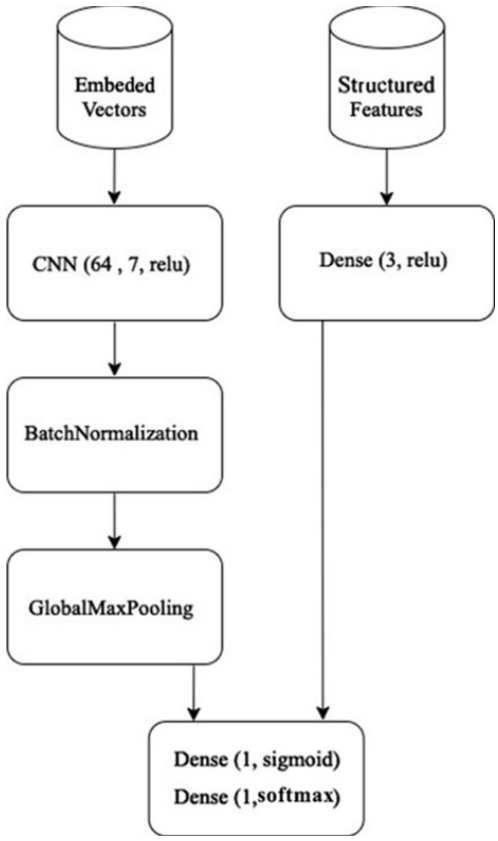
In deep learning, batch size affects training efficiency: large batches can move learning away from the optimal point, while small batches increase training time. In this study, batch sizes of 16, 32, and 64 were tested. Model evaluation used 5-fold cross-validation, with each data point serving as both training and validation data. Regularization, such as batch normalization, stabilized and accelerated training. Resampling techniques were applied within the training data to avoid data leakage. Finally, neural network architectures and feature combinations were optimized using a genetic algorithm based on cross-validated F1-score to improve generalization, enhancing the reliability of the loyalty prediction framework.

3.10 Model Evaluation

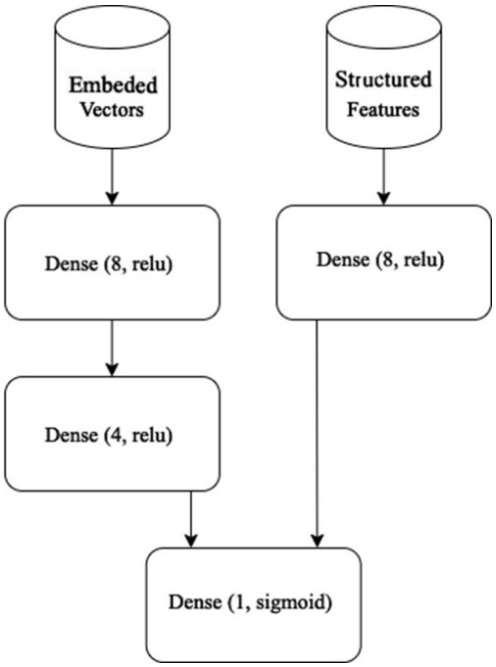
The resulting model in each scenario was evaluated using 5-fold cross-validation, and the F1-score metric was calculated according to Equations 1–3. The obtained value was used as the output of the optimization algorithm's fitness function. This metric was computed as a macro average for the three-class scenarios using Equations 4–6. This approach is useful for datasets in which each class has equal importance,



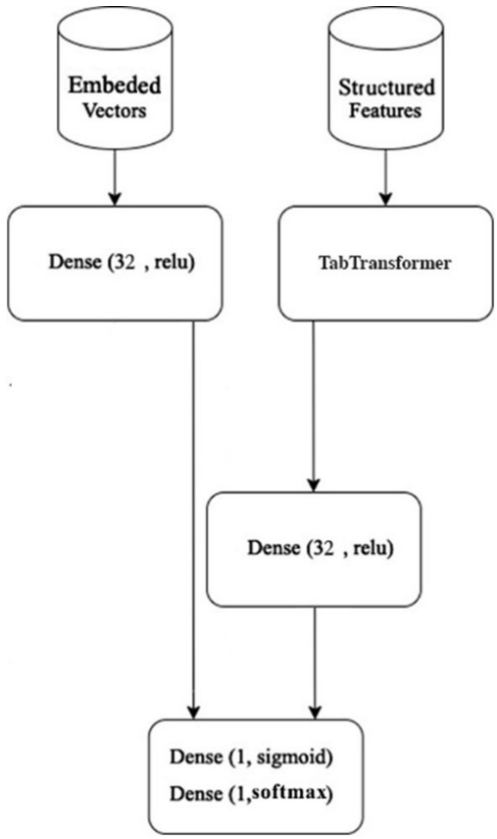
(a) Classification Architecture A.



(b) Classification Architecture B.



(c) Classification Architecture C.



(d) Classification Architecture D.

Figure 3. Architectures A, B, C, and D Applied for Loyalty Prediction.

as it assigns equal weight to all classes. The accuracy metric was also computed, as defined in Equation 7.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Macro Precision} = \frac{1}{N} \sum_{i=1}^{N} \text{Precision}_i \tag{4}$$

$$\text{Macro Recall} = \frac{1}{N} \sum_{i=1}^{N} \text{Recall}_i \tag{5}$$

$$\text{Macro F1-score} = \frac{1}{N} \sum_{i=1}^{N} F1\text{-score}_i \tag{6}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Here TP, TN, FP, FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively, and N denotes the number of classes.

3.11 Optimization

In this study, a genetic algorithm is employed with the primary objective of efficiently identifying the optimal combination of model parameters for customer loyalty prediction. Given the large and complex search space (19,440 possible configurations derived from combinations of language models, RFM settings, feature sets, neural network architectures, and training parameters), exhaustive evaluation is computationally expensive and impractical. Therefore, the genetic algorithm is used as an intelligent search strategy to explore this space and converge toward high-performing solutions by iteratively improving candidate models based on their predictive performance (F1-score). This enables the selection of a near-optimal model configuration while significantly reducing computational cost.

The settings for the genetic algorithm were chosen to include 50 generations, each with 10 chromosomes. Each chromosome represents a possible combination of parameters and is essentially a potential solution for constructing and training the customer loyalty prediction models. Consequently, 50 generations, each containing 10 solutions, are produced and evaluated, resulting in a total of 500 different solutions assessed. This number is significantly smaller compared to evaluating the entire solution space. The details of the genetic algorithm used in this study will be described further.

Representation: In the genetic algorithm used in this study, each chromosome is represented as a one-dimensional array of parameter values. For example, a chromosome with genes represented as [1, 1, 5, 4, 3, 1, 2, 3, 1] corresponds to a configuration including the first language model (BERT), the first set of structured features (title and actual review scores), RFM settings as R=5, F=4, M=3, consideration of two different labels for the problem, a under-sampling method, a batch size of 64, and finally, a neural network with the first type of architecture. Table 2 displays the values of the genes for the chromosomes defined for the genetic algorithm.

Initial Population: Consists of 10 individuals randomly selected. Fitness Function: In the genetic algorithm used in this study, the fitness function calculates the F1-score for each different combination of model construction and training (a chromosome) using 5-fold cross-validation.

Parent Selection: Parents are selected in each generation by randomly choosing six chromosomes from the current population 10 times (equal to the chromosome count). The best chromosomes based on the fitness function are then selected as parents for the next generation, resulting in 10 parents for that stage.

Mutation: Mutation in the genetic algorithm occurs in 20% of cases. One gene of the offspring is randomly selected for mutation (e.g., the gene related to the language model) and then assigned a random value from the permitted values for that gene.

Crossover: In this algorithm, crossover is performed by selecting pairs of parents sequentially. A random crossover point is chosen within the gene structure of a chromosome. Genes before this random point are transferred from the first parent to the first offspring and from the second parent to the second offspring. Genes after the random point are transferred from the first parent to the second offspring and from the second parent to the first offspring. Thus, each pair of parents produces two new offspring.

Next Generation Creation: After crossover and mutation based on the selected parents, the resulting offspring form the population for the next generation.

3.12 Model Explanations

SHAP is a model-agnostic interpretability method that decomposes each prediction into additive contributions from input features [43]. In our case, each textual feature (word or phrase) in a customer review is treated as a "player" whose contribution to the loy-

Table 2. Optimization Parameters.

Parameter	Values	Description
LM (Language Model)	1: BERT 2: ALBERT2 3: RoBERTa 4: GPT-2 5: Qwen 1.5-0.5B	The language model used for embedding customer review texts
SD (Structured Data)	1: Title Presence and Actual Review Score 2: Title Presence and Predicted Review Score 3: Actual Review Score	Structured features combined with vectors obtained from language models
R	3 , 4 , 5	Recency value in RFM model
F	3 , 4 , 5	Frequency value in RFM model
M	3 , 4 , 5	Monetary value in RFM model
DL (Distinct Labels)	1: Two-class 2: Three-class	Loyalty labels types
DB (Dataset Balancing)	1: Oversampling 2: Undersampling	Dataset balancing method
BS (Batch Size)	16, 32, 64	The number of training samples in each batch during the training of the loyalty prediction model
ML (Machine Learning)	1: Classification architecture A (CNN(6, 3)) 2: Classification architecture B (CNN(64, 7)) 3: Classification architecture C (Dense)	Deep learning architecture employed

alty score is quantified by a SHAP value. In practice, SHAP runs on the trained deep learning model to produce these values: the predicted loyalty score is broken down into the sum of feature contributions. SHAP provides local explanations by attributing an individual prediction to its input features, and global explanations by aggregating these attributions over the dataset.

SHAP values can be computed for each word in a review to visualize its impact on loyalty prediction. Red-highlighted words increase the loyalty score, while blue ones decrease it. For example, “excellent service” might be red (positive), and “wait” blue (negative). This provides local explanations for individual predictions. Globally, averaging SHAP values across reviews ranks words by their overall impact on loyalty. For instance, terms like “quality” and “support” may drive loyalty, while “delay” and “refund” indicate disloyalty. SHAP enables businesses and regulators to understand and verify the model’s behavior by identifying key review features, such as “fast shipping,” that influence loyalty.

4 Results

4.1 Dataset description

The dataset used in this study contains information on orders placed by customers of a Brazilian e-commerce platform called Olist, with orders spanning from 2016 to 2018. The original language of the dataset is Brazilian Portuguese, which was translated into English using Google’s machine translation system. A valid and comprehensive dataset in English was not available at the time of selecting a suitable dataset for this study.

On the Olist e-commerce platform, after customers complete and receive their orders, they are asked to fill out a questionnaire. This questionnaire captures the customers’ textual feedback (including the title and content of the review) as well as their numeri-

cal score of the completed order. The dataset also includes the necessary information for calculating customer behavioral loyalty based on the RFM model. Specifically, the F parameter was calculated for each customer based on the number of orders delivered to them, the R parameter was determined based on the date of the most recent order delivered, and the M parameter was derived from the total monetary value of the orders delivered to each customer. For the RFM scoring, only successfully delivered orders were considered; canceled or failed orders were excluded. Additionally, for building predictive models, only orders with non-empty review texts were included. The final number of data points used to build and train the customer loyalty prediction model was 58,841 samples. Figure 4 presents the distribution of features utilized for loyalty prediction based on the Olist dataset.

4.2 Sentiment Analysis Results

As previously mentioned, in order to use predicted scores instead of actual scores in the structured features, models were built and compared to predict these numerical scores, and the best model was selected for this purpose. To identify the suitable model, all ten models, including two neural network architectures and five language models for text vectorization, were executed separately, and the R^2 score was calculated for them using 5-fold cross-validation.

The results of these executions are shown as a bar chart in Figure 5. According to this chart, the R^2 score for the model built using the first architecture and the ALBERT language model was 0.81, which was the best result. Therefore, this model was used to predict the numerical score of the reviews.

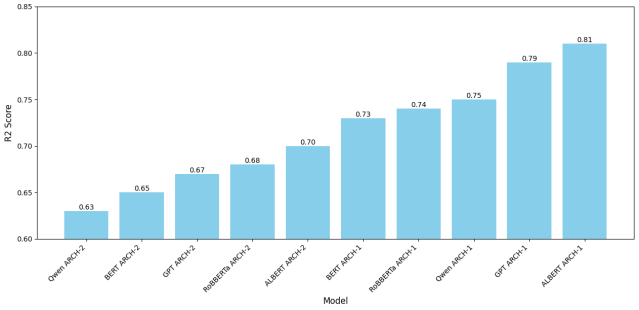


Figure 5. Evaluation Results of Sentiment Analysis Models.

4.3 Optimization Results

After running the genetic optimization algorithm to evaluate different parameter values for building customer loyalty prediction models, the optimal parameters were determined, as shown in Table 3. Table 4 illustrates the evaluation results of the optimal model obtained through the genetic algorithm, assessed using 5-fold cross-validation.

Table 3. Optimum values of parameters obtained by genetic algorithm for proposed method.

Parameter	Optimum value
R	3
F	5
M	5
LM	BERT
BS	64
DB	Oversampling
ML	TabTransformer
SD	Title Presence & Actual Review Score
DL	Two-class

Table 4. Evaluation results of the optimum model for loyalty prediction.

Criteria	Value
F1-score	90.11
Recall	88.41
Precision	91.88
Accuracy	92.03

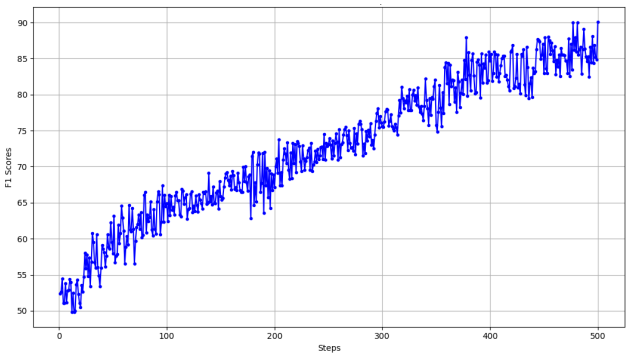
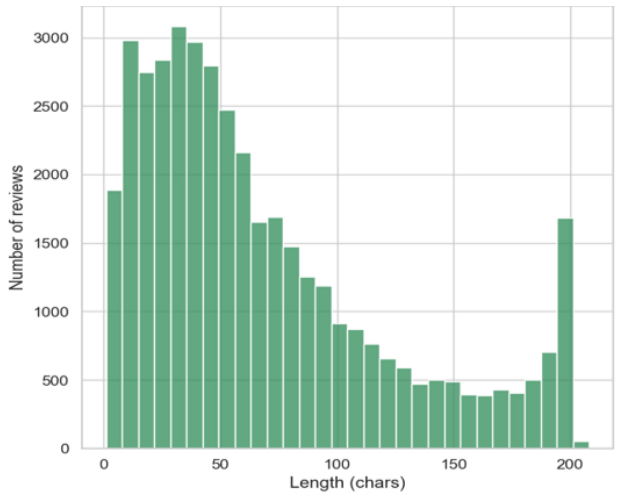


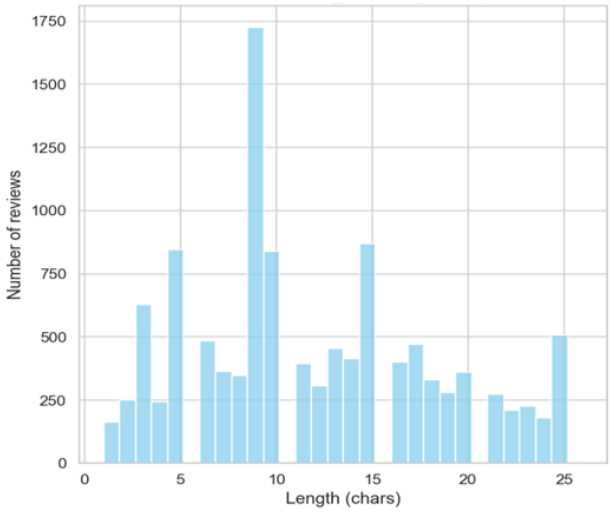
Figure 6. The changes of F1-score for each of the 500 iterations of training the predictive mode.

In this study, each neural network was trained for 200 epochs. To prevent overfitting, early stopping was applied if validation loss didn't improve for 25 epochs. The training data was split into training and test sets in 5-fold cross-validation, resulting in each model being trained in five stages. The final performance of the neural network was calculated as the average performance across all five stages.

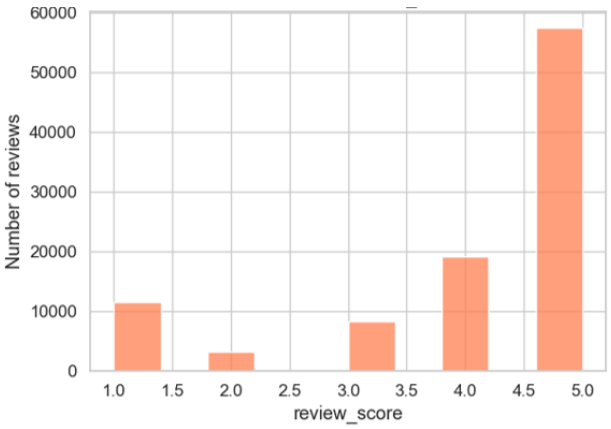
Since the settings used in the genetic algorithm for this problem are configured with 50 generations,



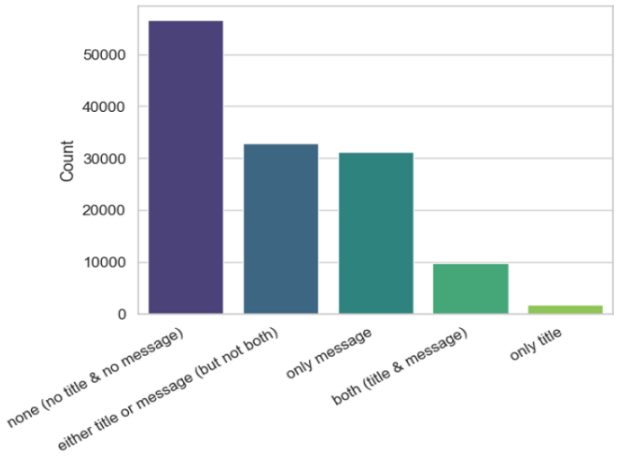
(a) Distribution of Review Message Length.



(b) Distribution of Review Title Length.



(c) Distribution of Review Score.



(d) Distribution of Reviews Regarding Presence of Title/Message.

Figure 4. Distribution of review features used in loyalty prediction: (a) review message length, (b) review title length, (c) review score, and (d) number of reviews regarding the presence of title/message.

each consisting of a population of 10 chromosomes (10 potential solutions for constructing and training predictive models), a total of 500 different solutions are evaluated in this algorithm. The trend of the F1-score across each of these 500 evaluations is shown in Figure 6.

For a deeper analysis and evaluation, the results of the best model's assessment are shown in Figure 7, with certain parameters frozen according to the scenarios described in Table 5.

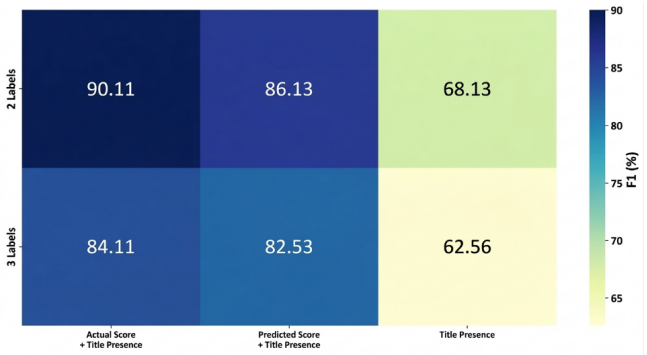


Figure 7. Evaluation results (F1-score) for the best models across different scenarios.

Table 5. Evaluation Scenarios.

Scenario	Structured features	# of Labels
1	Actual Review Score Title Presence	2
2	Predicted Review Score Title Presence	2
3	Actual Review Score	2
4	Actual Review Score Title Presence	3
5	Predicted Review Score Title Presence	3
6	Actual Review Score	3

4.4 Comparison of the Proposed Method and Baseline Models

To assess the effectiveness of our proposed loyalty prediction framework, we conducted comparative experiments against a series of baseline configurations that combine non-transformer text vectorization models with classical machine learning algorithms. Non-transformer text vectorization methods employed are as follows.

GloVe: A foundational word vectorization algorithm that creates word embeddings by utilizing global co-occurrence statistics, providing robust and meaningful representations for a wide range of NLP tasks.

fastText: An efficient word embedding method that leverages subword information to create robust word representations, making it especially useful for languages with complex word formation and for handling noisy or user-generated text.

Word2Vec: A foundational technique in NLP for generating word embeddings, enabling machines to understand and process human language by mapping words to a numerical space that preserves their contextual and semantic relationships.

Classical machine learning algorithms employed in baseline methods are as follows.

XGBoost (eXtreme Gradient Boosting): A scalable and flexible tree boosting system that achieves high predictive performance, particularly on tabular datasets, and is a go-to method for many practical machine learning challenges.

SVM (Support Vector Machine): A supervised learning algorithm that finds the best boundary (hyperplane) to separate classes, using only the most critical data points (support vectors), and can be adapted to handle non-linear data through kernel methods. RF (Random Forest): A flexible and powerful machine learning algorithm that builds an ensemble of decision trees to achieve high accuracy and generalization, making it a popular choice for both classification and regression problems.

Table 6. Optimum values of parameters obtained by genetic algorithm for baseline method.

Parameter	Optimum value (binary/three class)
R	3,4
F	4,5
M	5,5
LM	fastText,GloVe
BS	64,32
DB	Oversampling,Oversampling
ML	RF,RF
SD	Title Presence and Actual Review Score

It is noteworthy that, regarding the baseline settings, a genetic algorithm with configurations identical to those employed in the proposed method was applied to both binary and three-class classification tasks utilizing non-transformer text vectorization techniques (GloVe, FastText, Word2Vec) and classical machine learning models (XGBoost, SVM, Random Forest (RF)). The optimal hyperparameter configurations derived for each of these two experimental setups are summarized in Table 6.

The experimental results demonstrate that the proposed method achieved a better F1-score (Figure 8).

4.5 Feature Importance Analysis

To better understand which features most significantly influenced the performance of our customer loyalty prediction model, we conducted a feature ablation study. In this analysis, we systematically removed one feature at a time from the input space and retrained the model to observe the impact on the F1-score. Each time, we retrained and tested the best-performing model by excluding one feature at a time. This approach helped us quantify the marginal contribution of

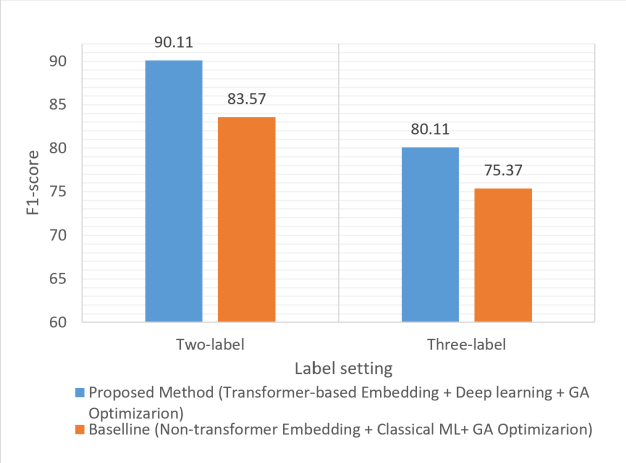


Figure 8. Comparison of the proposed and baseline loyalty prediction methods.

each feature to the final model's predictive capability. The results of this analysis are summarized in Table 7.

Table 7. The impact of features on loyalty prediction performance.

Best model	F1 Score	Percentage of decline
All	90.11	0.0
All without review text	56.9	-36.85
All without title presence	81.27	-9.81
All without review rating	83.44	-7.44

The review text alone contributes the most to model performance. Removing it resulted in a drastic 36.85% drop in F1-score. This demonstrates the model's heavy reliance on the contextual and semantic richness encoded within customer reviews. Words and phrases carry significant emotional and behavioral cues that help the model infer loyalty, especially when structured signals are ambiguous or weak. This finding supports our use of advanced transformer-based models (e.g., BERT, GPT, Qwen), which can effectively capture these subtle textual patterns.

Excluding the binary indicator for title presence led to a noticeable 9.81% drop in performance. While seemingly a simple feature, the presence of a review title may correlate with the customer's level of engagement or intent. Our statistical analysis found that loyal customers were 11.37 times more likely to include a title in their reviews compared to non-loyal customers. This aligns with the hypothesis that more loyal users are more expressive, detailed, and deliberate in their feedback (traits captured in this behavioral indicator).

Although numerical review ratings (1 to 5 stars) are widely used in e-commerce, removing this structured feature only reduced the F1-score by 7.44%, indicating that while helpful, ratings are not as powerful in isolation. This may be due to the inherent noise in star ratings (for example, some customers may leave a 4-star rating despite providing very positive textual feedback, or vice versa). Thus, text-based sentiment often provides richer behavioral signals than the rating alone.

The performance gap between "all features" and "all without review text" suggests that structured features (review rating and title presence) alone are insufficient to robustly model loyalty. However, their inclusion with text embeddings improves performance, suggesting synergistic effects. The model benefits from both the semantic nuance of text and the structured regularity of numeric data, especially in edge cases where one modality may be ambiguous. These insights are critical for real-world deployment. For instance, in environments where structured data (e.g., ratings) is unavailable (such as social networks or third-party review platforms), the model still performs well using text alone. Conversely, when textual data is limited (e.g., very short reviews), structured signals can partially compensate.

4.6 Explanations

Figures 9-11 illustrate three distinct examples of local explanations generated by SHAP. In these figures, red highlights features with higher values (e.g., elevated numerical measures or the presence of a categorical attribute), which amplify the model's predicted output. Conversely, blue denotes lower feature values (e.g., reduced numerical quantities or the absence of a categorical feature), thereby diminishing the prediction. This color-coding illustrates how each feature's magnitude or status directly influences the model's final decision.

Example 1: Loyal

Figure 9. SHAP Explanation Example 1.

Example 2: Non loyal

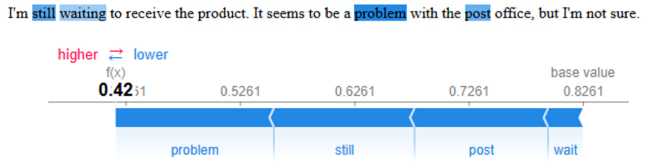


Figure 10. SHAP Explanation Example 2.

Example 3: Non loyal

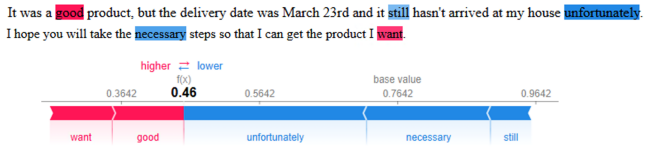


Figure 11. SHAP Explanation Example 3.

Across Figures 12–14, the length and hue of each bar convey how individual words drive the model's loyalty prediction: red bars indicate terms whose presence raises the predicted loyalty score, whereas blue bars indicate those that lower it. In Figure 12 (global view over 500 reviews), the top 30 words by absolute SHAP magnitude reveal which terms the model relies on most heavily. For example, words like "recommend", "great", "clarify", and "reliable" appear as long red bars, indicating they consistently elevate loyalty predictions when present; conversely, terms such as "unfortunately", "delay", "bad", and "stock" show up as long blue bars, signifying they regularly suppress loyalty scores.

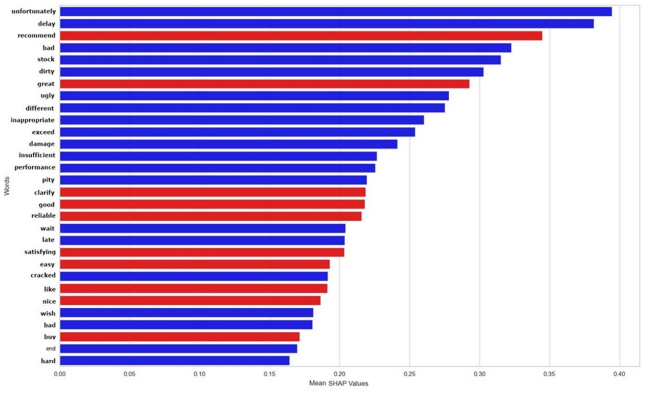


Figure 12. SHAP Global Explanation for 500 Test Data.

Figure 13 isolates only those reviews that the model labeled "Loyal." Here, the red bars dominate among the top 20 words, underscoring that positive sentiment markers drive these loyalty classifications. Any shorter blue bars illustrate that even in overall positive reviews, occasional negative references slightly temper, but do not overturn, the loyalty prediction. By contrast, Figure 14 focuses on "Non loyal" instances. The top 20 words are predominantly blue, highlighting that negative descriptors chiefly determine non-loyalty labels. Red bars indicate mildly positive words that the model found less persuasive in shifting a non-loyal classification toward loyalty.

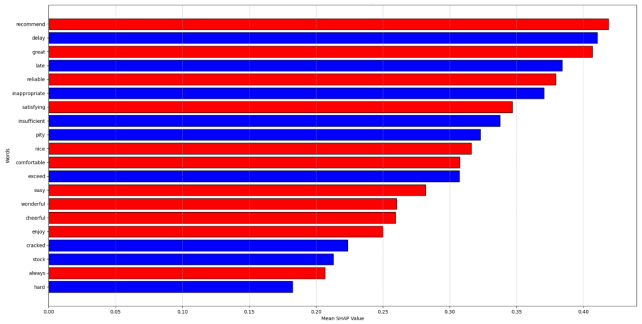


Figure 13. SHAP Global Explanation for "Loyal" Instances.

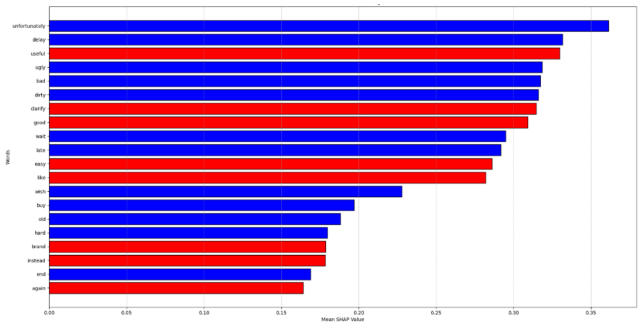


Figure 14. SHAP Global Explanation for "Non loyal" Instances.

4.7 Discussion

The results of this study demonstrate that customer behavioral loyalty can be predicted with a high degree of accuracy using customer review data, even in the absence of traditional transactional indicators. Across both binary and three-class classification settings, the proposed framework consistently achieved strong performance, confirming the effectiveness of combining deep learning architectures with transformer-based language models for extracting meaningful behavioral signals from textual data.

A key finding of this study is the relatively small performance gap between models trained using actual review scores and those relying on predicted scores. Although models incorporating true numerical ratings achieved slightly higher F1-scores, the decrease in performance when using predicted scores remained limited in both classification settings. This indicates that explicit rating information, while beneficial, is not strictly necessary for accurate loyalty prediction.

Instead, the semantic and contextual information embedded in review texts appears to capture much of the underlying signal associated with customer loyalty. From a practical perspective, this finding is particularly important, as it demonstrates that the proposed approach remains applicable in environments where numerical ratings are unavailable, incomplete, or unreliable.

Another important observation concerns the role of the "title presence" feature. Excluding this feature resulted in a noticeable decline in model performance, highlighting its predictive value. This suggests that seemingly simple behavioral indicators can provide meaningful insights into customer engagement. Supporting this, statistical analysis revealed that loyal customers are significantly more likely to include titles in their reviews compared to non-loyal customers. This behavior may reflect a higher level of involvement, effort, or expressiveness, which in turn correlates with stronger loyalty. The result underscores the importance of incorporating lightweight yet informative structured features alongside text embeddings.

The findings also reinforce the dominant contribution of review text to model performance. As shown in the feature ablation analysis, removing textual input led to a substantial degradation in predictive accuracy, far exceeding the impact of removing structured features such as ratings or title presence. This highlights the richness of unstructured textual data in capturing nuanced aspects of customer experience, including sentiment, intent, and implicit behavioral cues. At the same time, the improvement observed when combining text with structured features suggests a complementary relationship between these data modalities, where each contributes distinct but synergistic information.

An additional practical consideration is the model's robustness in scenarios with limited customer interaction history. As depicted in Figure 15, the dataset analysis showed that most customers provided only a single review, yet the model maintained strong predictive performance. This indicates that the approach is capable of inferring loyalty from minimal input, making it suitable for early-stage customer evaluation, new user analysis, and cold-start scenarios where historical behavioral data is sparse or unavailable.

Although the proposed model is developed and evaluated using data from a single e-commerce platform (Olist), its design is not platform-specific. The framework relies on two main components: customer review text and behavioral loyalty labels derived from the RFM model. Since customer reviews are widely available across various domains, including retail, hospitality, telecommunications, and online services, and RFM-based behavioral indicators can be constructed

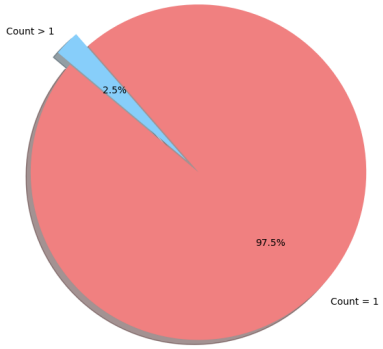


Figure 15. Distribution of Review Counts per Customer.

wherever transactional data exists, the proposed approach can be adapted to different industries. Moreover, the use of transformer-based language models enhances the model's ability to generalize across domains, as these models are pre-trained on diverse corpora and capture broad linguistic patterns. However, domain-specific fine-tuning and recalibration of RFM parameters may be necessary to achieve optimal performance in different markets or industries.

5 Conclusions

In the competitive landscape of virtual businesses, the importance of customer loyalty has become increasingly evident. Additionally, with the expansion of user presence on the internet, a vast amount of unstructured data has been generated, encompassing extensive information on various topics, including customer loyalty. Despite this, there is a lack of studies in the field of customer loyalty prediction that leverage the potential of this unstructured data. This study addresses this gap by developing a model that predicts behavioral customer loyalty (based on the RFM model) through the analysis of customer review texts from an online store. This represents an innovative aspect of the study. In the best-case scenario, the model for predicting customer loyalty achieved an accuracy of 0.90 for binary classification and 0.80 for three-class classification. Therefore, by analyzing customer review texts and utilizing neural networks and natural language processing techniques, it is possible to predict customer behavioral loyalty with a high degree of accuracy.

The feature importance analysis not only confirms the superiority of transformer-based textual embeddings but also highlights the practical value of combining them with simple structured features. It also validates our model architecture's design decision to process structured and unstructured inputs through separate neural branches, ensuring both modalities contribute meaningfully without being overshadowed.

While the model's predictive accuracy is technically impressive, its real-world value becomes most apparent when considering how businesses can leverage it to improve customer retention. In many e-commerce settings, behavioral loyalty is typically assessed through retrospective transactional data, which may not be available for new customers or competitors' customers. This study presents an innovative alternative, predicting loyalty based solely on customer reviews, which provides timely and actionable insights even in the absence of long purchase histories. By integrating this model into a company's CRM system, businesses can monitor incoming customer reviews in real-time and immediately assess each customer's likelihood of continued engagement. For instance, if the model flags a customer as potentially non-loyal based on the tone, sentiment, or content of their review, the business can trigger targeted interventions such as personalized offers, loyalty rewards, or direct follow-up communication to address dissatisfaction and prevent churn. This makes it possible to move from reactive to proactive retention strategies.

Additionally, because the model includes an explainability layer using SHAP, it does not function as a black box. Instead, it identifies which textual elements most strongly influenced the loyalty prediction. This interpretability allows marketing and customer service teams to extract granular insights into what drives loyalty or dissatisfaction. For example, if SHAP highlights that negative comments about delivery time are strongly associated with predicted disloyalty, the company can focus on improving logistics or better setting delivery expectations.

Despite the promising results, this study has several limitations that should be acknowledged. First, the proposed model relies heavily on the availability of sufficiently large volumes of customer review text. In scenarios where reviews are sparse, extremely short, or absent, the model's predictive performance may degrade, as the textual component plays a central role in capturing behavioral signals. Second, the quality of the predicted loyalty labels is inherently dependent on the initial RFM-based segmentation. Any noise, bias, or suboptimal parameter selection in the RFM labeling process may propagate into the training data and affect the model's accuracy. Furthermore, although the use of a genetic algorithm helps optimize RFM configurations, the labeling process still reflects an approximation of true customer loyalty rather than a ground-truth measure.

The model is trained and evaluated on data from a single e-commerce platform, which may limit its generalizability across different industries, cultural contexts, or customer behavior patterns. While the methodology is transferable, domain-specific adaptation and validation are required. Moreover, the approach depends on the quality of textual data, which may include noise such as informal language, grammatical errors, sarcasm, or translation artifacts. These factors can reduce the effectiveness of NLP models in accurately capturing sentiment and intent.

The proposed direction for future research is to enhance the current method of customer loyalty labeling by implementing a combined approach. This would involve labeling customer loyalty in a manner that considers both behavioral and attitudinal loyalty, thus covering the limitations and challenges of each type of loyalty.

The current segmentation approach is based solely on the traditional RFM model. Exploring alternative customer segmentation techniques, particularly more recent methods that leverage machine learning, could further enhance the accuracy and effectiveness of the predictive model.

Additionally, extending the model to incorporate temporal trends in customer scores or other dynamic features could enhance loyalty prediction. Analyzing how changes in customer attributes over time influence loyalty outcomes may yield deeper insights and improve predictive accuracy. This direction aligns with the growing need for adaptive models in customer relationship management.

References

[1] Adnan Aktepe, Süleyman Ersöz, and Bilal Toklu. Customer satisfaction and loyalty analysis with classification algorithms and structural equation modeling. *Computers & Industrial Engineering*, 86:95–106, 2015. doi:10.1016/j.cie.2014.09.031. Applications of Computational Intelligence and Fuzzy Logic to Manufacturing and Service Systems.

[2] Jo-Ting Wei, Shih-Yen Lin, and Hsin-Hung Wu. A review of the application of rfm model. *African journal of business management*, 4(19):4199, 2010.

[3] Zeynep Hilal Kilimci. Prediction of user loyalty in mobile applications using deep contextualized word representations. *Journal of Information and Telecommunication*, 6(1):43–62, 2022. doi:10.1080/24751839.2021.1981684.

[4] Wissam Nazeer Wassouf, Ramez Alkhatib, Kamal Salloum, and Shadi Balloul. Predictive analytics using big data for increased customer loyalty: Syriatel telecom company case study. *Journal of Big Data*, 7(1):29, 2020. doi:10.1186/s40537-020-00290-0.

[5] Jesselin Vemberain and Abdullah Rakhman. Influence of promotions, price perceptions, service quality towards customer loyalty through customer satisfaction gojek in jakarta. *Jurnal Ekonomi Perusahaan*, 31(1):21–40, 2024. doi:10.46806/jep.v31i1.1108.

[6] Shih-I Cheng. Comparisons of competing models between attitudinal loyalty and behavioral loyalty. *International Journal of Business and Social Science*, 2(10):149–166, 2011.

[7] Nikhil Patel and Sandeep Trivedi. Leveraging predictive modeling, machine learning personalization, nlp customer support, and ai chatbots to increase customer loyalty. *Empirical Quests for Management Essences*, 3(3):1–24, 2020.

[8] Chih-Hsuan Wang. Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. *Expert systems with applications*, 37(12):8395–8400, 2010. doi:10.1016/j.eswa.2010.05.042.

[9] Myint Zaw and Pichaya Tandayya. Product categorization for social marketing applying the rfc model and data science techniques. *International Journal of Business Analytics (IJBAN)*, 7(4):43–62, 2020. doi:10.4018/IJBAN.2020100104.

[10] Zeyad M Kishada, Norailis AB Wahab, and Aouache Mustapha. Customer loyalty assessment in malaysian islamic banking using artificial intelligence. *Journal of Theoretical & Applied Information Technology*, 87(1), 2016.

[11] Hossein Alizadeh and B Minaei Bidgoli. Introducing a hybrid data mining model to evaluate customer loyalty. *Engineering, Technology & Applied Science Research*, 6(6):1235–1240, 2016. doi:10.48084/etasr.741.

[12] Marcos Roberto Machado, Salma Karray, and Ivaldo Tributino De Sousa. Lightgbm: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In *2019 14th international conference on computer science & education (ICCSE)*, pages 1111–1116. IEEE, 2019. doi:10.1109/ICCSE.2019.8845529.

[13] T Rahul and R Majhi. An adaptive nonlinear approach for estimation of consumer satisfaction and loyalty in mobile phone sector of india. *Journal of Retailing and Consumer Services*, 21(4):570–580, 2014. doi:10.1016/j.jretconser.2014.03.009.

[14] Heni Sulistiani, Kurnia Muludi, and Admi Syarif. Implementation of dynamic mutual information and support vector machine for customer loyalty classification. In *Journal of Physics: Conference Series*, volume 1338, page 012050. IOP Publishing, 2019. doi:10.1088/1742-6596/1338/1/012050.

[15] Cristián J Figueroa. Predicting customer loyalty labels in a large retail database: A case study in chile. In *Data Mining: Special Issue in Annals of Information Systems*, pages 229–253. Springer, 2009. doi:10.1007/978-1-4419-1280-0_10.

[16] Anida Granov. Customer loyalty, return and churn prediction through machine learning methods: for a swedish fashion and e-commerce company, 2021.

[17] Hiu Fai Lee and Ming Jiang. A hybrid machine learning approach for customer loyalty prediction. In *International Conference on Neural Computing for Advanced Applications*, pages 211–226. Springer, 2021. doi:10.1007/978-981-16-5188-5_16.

[18] Mohamed M Abbassy. Using machine learning technique for analytical customer loyalty. *International Journal of Computer Science & Network Security*, 23(8):190–198, 2023.

[19] Xujie Qin. Research on loyalty prediction of e-commerce customer based on data mining. *Applied Mathematics and Nonlinear Sciences*, 8(2):721–732, 2023. doi:10.2478/amns.2022.1.00020.

[20] Heni Sulistiani and Aris Tjahyanto. Comparative analysis of feature selection method to predict customer loyalty. *IPTEK The Journal of Engineering*, 3(1), 2017. doi:10.12962/j23378557.v3i1.a2257.

[21] Andri Wijaya and Abba Suganda Girsang. Use of data mining for prediction of customer loyalty. *Communication and Information Technology Journal*, 10(1):41–47, 2016. doi:10.21512/commit.v10i1.1660.

[22] Abdul Syukur, Romi Satria Wahono, Abdul Razak Naufal, and Catur Supriyanto. Bootstrapping and weighted information gain in support vector machine for customer loyalty prediction. *Journal of Internet Banking and Commerce*, 23(1):1–13, 2018.

[23] Lili Tong, Yiting Wang, Fan Wen, and Xiaowen Li. The research of customer loyalty improvement in telecom industry based on nps data mining. *China Communications*, 14(11):260–268, 2017. doi:10.1109/CC.2017.8233665.

[24] Iskandar Zul Putera Hamdan and Muhaini Othman. Predicting customer loyalty using machine learning for hotel industry. *Journal of Soft Computing and Data Mining*, 3(2):31–42, 2022. doi:10.30880/jscdm.2022.03.02.004.

[25] Chi-I Hsu, Meng-Long Shih, Biing-Wen Huang, Bing-Yi Lin, and Chun-Nan Lin. Predicting tourism loyalty using an integrated bayesian network mechanism. *Expert Systems with Applications*, 36(9):11760–11763, 2009. doi:10.1016/j.eswa.2009.04.010.

[26] Iskandar Zul Putera Hamdan, Muhaini Othman, Yana Mazwin Mohmad Hassim, Suziyanti Marjudi, and Munirah Mohd Yusof. Customer loy-

alty prediction for hotel industry using machine learning approach. *JOIV: International Journal on Informatics Visualization*, 7(3):695–703, 2023. doi:10.30630/joiv.7.3.1335.

[27] Balgopal Singh. Predicting airline passengers' loyalty using artificial neural network theory. *Journal of air transport management*, 94:102080, 2021. doi:10.1016/j.jairtraman.2021.102080.

[28] Jain-Shing Wu, Ting-Hsuan Chien, Li-Ren Chien, and Chin-Yi Yang. Using artificial intelligence to predict class loyalty and plagiarism in students in an online blended programming course during the covid-19 pandemic. *Electronics*, 10(18):2203, 2021. doi:10.3390/electronics10182203.

[29] Chuang Wang, Rongxin Zhou, and Matthew KO Lee. Can loyalty be pursued and achieved? an extended rfd model to understand and predict user loyalty to mobile apps. *Journal of the Association for Information Science and Technology*, 72(7):824–838, 2021. doi:10.1002/asi.24448.

[30] Fakhroddin Noorbehbahani, Soroush Bajoghli, and Hooman Hoghooghi Esfahani. Customer loyalty prediction of e-marketplaces via review analysis. In *2023 9th International Conference on Web Research (ICWR)*, pages 311–316. IEEE, 2023. doi:10.1109/ICWR57742.2023.10139037.

[31] Usman Ghani, Imran Sarwar Bajwa, and Aimen Ashfaq. A fuzzy logic based intelligent system for measuring customer loyalty and decision making. *Symmetry*, 10(12):761, 2018. doi:10.3390/sym10120761.

[32] Rumana Shahid, Md Abu Sufian Mozumder, Md Murshid Reja Sweet, Mehedi Hasan, Mahfuz Alam, Mohammad Anisur Rahman, Mani Prabha, Md Arif, Md Parvez Ahmed, and Md Rafiqul Islam. Predicting customer loyalty in the airline industry: a machine learning approach integrating sentiment analysis and user experience. *International Journal on Computational Engineering*, 1(2):50–54, 2024. doi:10.62527/comien.1.2.12.

[33] Rida Khan and Siddhaling Urolagin. Airline sentiment visualization, consumer loyalty measurement and prediction using twitter data. *International journal of advanced computer science and applications*, 9(6), 2018. doi:10.14569/IJACSA.2018.090652.

[34] Siddhaling Urolagin and Saifali Patel. User-specific loyalty measure and prediction using deep neural network from twitter data. *IEEE Transactions on Computational Social Systems*, 11(1):1046–1061, 2023. doi:10.1109/TCSS.2023.3239523.

[35] Mohamed Zaki, Dalia Kandeil, Andy Neely, and Janet R McColl-Kennedy. The fallacy of the net promoter score: Customer loyalty predictive model. *Cambridge Service Alliance*, 10(1):1–25, 2016.

[36] Mariana Avelino, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. A test suite for the evaluation of portuguese-english machine translation. In *International Conference on Computational Processing of the Portuguese Language*, pages 15–25. Springer, 2022. doi:10.1007/978-3-030-98305-5_2.

[37] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995. doi:10.1145/219717.219748.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[39] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

[41] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[42] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020. doi:10.48550/arXiv.2012.06678.

[43] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pages 4593–4603, 2022.

Pedram Kazemi Esfe received both his B.Sc. and M.Sc. degrees in Information Technology from the University of Isfahan, where his master's studies focused on E-commerce. His primary research interests include Natural Language Processing, Machine Learning, Data Science, and Customer Analytics.

Fakhroddin Noorbehbahani is an assistant professor of computer engineering at the University of Isfahan. He received his B.Sc. and M.Sc. in computer and IT engineering from, respectively, Isfahan University of Technology and Amirkabir University of Technology, Iran. He obtained his Ph.D. degree in computer engineering from Isfahan University of Technology, Iran. His research interests include Human-Computer Interaction, E-business, Machine Learning, and Data Science.