Journal of
*Theoretical and Applied Electronic Commerce Research*

MDPI

*Article*

# Optimizing Sentiment Analysis Models for Customer Support: Methodology and Case Study in the Portuguese Retail Sector

**Catarina Almeida [1], Cecilia Castro [2], Víctor Leiva [3,*], Ana Cristina Braga [4] and Ana Freitas [5]**

[1] Departament of Mathematics, Universidade do Minho, 4710-057 Braga, Portugal; pg46717@alunos.uminho.pt
[2] Centre of Mathematics, Universidade do Minho, 4710-057 Braga, Portugal; cecilia@math.uminho.pt
[3] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile
[4] ALGORITMI Research Centre, LASI, Universidade do Minho, 4710-057 Braga, Portugal; acb@dps.uminho.pt
[5] MC Sonae, Senhora da Hora, 4460-503 Matosinhos, Portugal; ancfreitas@mc.pt
* Correspondence: victor.leiva@pucv.cl or victorleivasanchez@gmail.com

**Abstract:** Sentiment analysis is a cornerstone of natural language processing. However, it presents formidable challenges due to the intricacies of lexical diversity, complex linguistic structures, and the subtleties of context dependence. This study introduces a bespoke and integrated approach to analyzing customer sentiment, with a particular emphasis on a case study in the Portuguese retail market. Capitalizing on the strengths of SentiLex-PT, a sentiment lexicon curated for the Portuguese language, and an array of sophisticated machine learning algorithms, this research constructs advanced models that encapsulate both lexical features and the subtleties of linguistic composition. A meticulous comparative analysis singles out multinomial logistic regression as the pre-eminent model for its applicability and accuracy within our case study. The findings of this analysis highlight the pivotal role that sentiment data play in strategic decision-making processes such as reputation management, strategic planning, and forecasting market trends within the retail sector. To the extent of our knowledge, this work is pioneering in its provision of a holistic sentiment analysis framework tailored to the Portuguese retail context, marking an advancement for both the academic field and industry application.

**Keywords:** consumer sentiment; lexical features; machine learning techniques; Python software; SentiLex-PT

check for updates

**Citation:** Almeida, C.; Castro, C.; Leiva, V.; Braga, A.C.; Freitas, A. Optimizing Sentiment Analysis Models for Customer Support: Methodology and Case Study in the Portuguese Retail Sector. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1493–1516. https://doi.org/10.3390/jtaer19020074

Academic Editor: Eduardo Álvarez-Miranda

Received: 4 January 2024
Revised: 21 May 2024
Accepted: 5 June 2024
Published: 10 June 2024

**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Sentiment analysis (SA) [1] is helpful in natural language processing (NLP), underpinning decision-making processes across various sectors such as commerce, finance, healthcare, and product analysis [2]. Machine learning (ML) [3] has been employed in various sectors for supporting the decision-making, including biomedical and healthcare data [4,5], data privacy [6], and forecasting [7].

Recent studies have focused on innovative data-driven methods for weighted SA [8,9]. The public sentiment for applications like customer SA, market prediction, and reputation management have been stated in [10–16]. However, SA faces obstacles such as informal writing, irony, language-specific nuances, and sarcasm, complicating sentiment detection and classification [17,18].

In light of these obstacles, investigations that break new ground by introducing pioneering, exhaustive, and hybrid methodologies for SA are needed, particularly in the Portuguese retail sector. The relevance of focusing on the retail sector is particularly heightened by the changes in customer behavior induced by the COVID-19 pandemic, an area that has received scholarly attention [19]. The efficacy of ML techniques has been proven in complex and dynamic scenarios, such as COVID-19 diagnosis [5,20], indicating their potential utility in analyzing the retail sector during these turbulent times.

In this context, we utilize SentiLex-PT, a specialized sentiment lexicon for the Portuguese language [21,22], to provide linguistic insights. These are then coupled with a range of ML techniques—including naive Bayes (NB), logistic regression (LR), multinomial logistic regression (MLR), ordinal regression (OR), support vector machines (SVMs), random forests (RFs), and eXtreme gradient boost (XGBoost)—to process complex data and offer a comprehensive analysis [20,23–25].

By coupling SentiLex-PT for sentiment extraction and ML models for sentiment classification, an effective approach to SA can be developed, especially when applied to Portuguese retail customer feedback. Recent research underscores the benefits of hybrid models. For instance, in [26], the authors employed Twitter data and reduced the feature count by up to 96% while outperforming most models. Such study evidenced the power of hybrid models with precision-tuned hyperparameters to transcend the performance of standalone models [27]. While their hybrid model yielded impressive results, they recognized that opportunities for enhancing performance remain through model tweaking and training. By adopting such a hybrid approach, a significant advance in SA can be obtained, amalgamating linguistic feature incorporation and sophisticated ML methods for more refined and precise sentiment classification. To the best of our knowledge, there is no methodology proposed until now considering the coupling of SentiLex-PT for sentiment extraction and ML for sentiment classification to produce hybrid models, especially in the Portuguese retail sector. However, ML has been widely used for various classification tasks such as cardiovascular diseases [28], abnormality detection [24], and even clustering [25].

The main objective of our investigation is to provide a novel and integrated methodology tailored for customer SA. This methodology incorporates various levels of analysis—document-level, sentence-level, and phrase-level—to capture the nuanced nature of sentiments. Document-level analysis yields an overarching sentiment score, while sentence-level dissection allows for contrasting sentiments within the same document.

On a more granular level, phrase-level SA identifies opinion words at the phrase-level, offering detailed scrutiny [29]. Thus, we provide a holistic understanding of sentiment by seamlessly integrating these levels of analysis. We apply this methodology to a case study in the Portuguese retail industry.

Beyond text pre-processing [30], our computational framework involves multiple steps. We start with feature extraction techniques to transform raw text into a machine-readable numerical format. We employ methods like bag-of-words, context capturing through N-grams, and manage sparse matrices through feature selection and reduction. Metrics such as precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC-AUC) are utilized for evaluation. Hyperparameter tuning is carried out using methods like grid search and randomized search [4,6]. Importantly, the selection of ML models was motivated by the specific nature of SA problems and by our dataset.

We deploy MLR for nominal sentiment classes, OLR for ordered sentiments, and SVMs for handling high-dimensional data. Multiscale monitoring with ML methods has also proven beneficial in other fields [23]. RFs are chosen for their robustness against overfitting and aptitude for managing large feature sets, while XGBoost is employed for its regularization capabilities, offering a balanced and high-performing model. The merit of adopting a supervised ML approach—where algorithms learn from labeled data—lies in its ability to generalize from training data to unseen inputs, following an iterative process of prediction, feedback, and parameter adjustment. We evaluate model performance and decide when to halt the learning process to achieve an optimal model. ML, in this context, is not just a computational tool but also a lens through which sentiment can be understood beyond mere definitions, considering contextual data, sarcasm, and misused words, similar to how ML is used for evaluating the impact on academic performance [31,32].

The remainder of this article is organized as follows. Section 2 discusses the methodology and ML classifiers. In Section 3, we cover the dataset and analytical methods. In Section 4, an empirical evaluation of various ML approaches for SA is provided. We conclude the study and outline future research avenues in Section 5.

## 2. Methodology

In this section, we provide a summary of the classifiers utilized in our approach to SA. We focus on key models such as MLR and XGBoost, which play a crucial role in our analysis. Additionally, we include a concise overview of other relevant models to ensure a comprehensive understanding of the methodologies applied. Our aim is to present these models in a manner that is both theoretically sound and empirically relevant to practitioners and researchers in the field of SA.

### *2.1. Naive Bayes*

NB is a simple yet effective ML algorithm for classification. The algorithm is based on the Bayes theorem and assumes that all features (covariates) are independent of each other. Despite this oversimplification, NB classifiers work well in many real-world situations, including text classification and spam filtering [33–36]. Given the class target variable $Y$ and features $X_1, \ldots, X_k$, with observed values $x_1, \ldots, x_k$, the NB model assumes that, under the naivety assumption that all features are independent of each other, the conditional probability is written as

$$P(Y = y | X_1 = x_1, \ldots, X_k = x_k) = \frac{P(Y = y) \prod_{i=1}^{k} P(X_i = x_i | Y = y)}{P(X_1 = x_1, \ldots, X_k = x_k)},$$

where the denominator $P(X_1 = x_1, \ldots, X_k = x_k)$ is a constant given the input, so $P(Y = y | X_1 = x_1, \ldots, X_k = x_k)$ can be stated as proportional to $P(Y = y) \prod_{i=1}^{k} P(X_i = x_i | Y = y)$. We choose the class $y$ that maximizes $P(Y = y) \prod_{i=1}^{k} P(X_i = x_i | y)$, leading to

$$y = \arg\max_{y} P(Y = y) \prod_{i=1}^{k} P(X_i = x_i | y).$$

The NB model is easy to build, with no complicated parameter estimation which makes it helpful. Additionally, it can handle a large number of features and is unaffected by irrelevant features, making it versatile for handling complex classification problems.

### *2.2. Logistic Regression*

LR is a powerful and commonly used ML algorithm for binary classification problems. It is a statistical model that utilizes a logistic function to describe a binary target variable $Y$ [37]. Unlike linear regression, which outputs continuous values, LR transforms its output to return a probability value that can be mapped to two classes: positive and negative. Given a feature vector $X_1, \ldots, X_k$ and its observed (input) values $x_1, \ldots, x_k$, the LR model first calculates a linear combination of these input features as $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, where $\beta_0, \beta_1, \ldots, \beta_n$ are the model parameters and $z$ is the log-odds or logit function that represents the linear combination of the input features weighted by the corresponding model parameters. Next, the model maps the log-odds $z$ to the probability of the positive class using the sigmoid function given by

$$\widehat{p} = 1/(1 + \exp(-z)),$$

where the output $\widehat{p}$ is the estimated probability of the positive class. Training the model involves finding the parameter $\beta$ that minimizes the cost function, typically the log-loss function as

$$J(\beta) = -\frac{1}{n} \sum_{i=1}^{n} (y_i \log(\widehat{p}_i) + (1 - y_i) \log(1 - \widehat{p}_i)),$$

where $n$ is the sample size, $y_i$ is the actual class label (positive or negative), and $\widehat{p}_i$ is the predicted probability of the positive class for instance $i$. This cost function is convex, which allows optimization algorithms like gradient descent to find the global minimum effectively. LR is widely used due to its efficiency and simplicity.

Despite its name, LR is primarily used for classification tasks rather than regression tasks. It works well for linearly separable classes and when the feature space is linearly related to the log-odds of the response (target or dependent) variable. However, it may not be as effective with non-linear problems or problems where feature interactions are important, as it assumes independence among the features. For multi-class classification problems, the variant called MLR is used, which will be discussed in the subsequent section.

### *2.3. Multinomial Logistic Regression*

In the context of our study, where the target variable $Y$ (sentiment) has five levels—very negative (VN), negative (N), neutral (Neu), positive (P), and very positive (VP)—an MLR model is particularly helpful.

MLR extends LR when the target variable has more than two classes, creating distinct LR models for each class against a reference, with its own regression coefficients. The key assumption, known as independence of irrelevant alternatives, posits that the selection of one class does not relate to the selection of any other. Given a categorical variable $Y$ with levels (VN, N, Neu, P, VP), $k$ covariates $\boldsymbol{X} = X_1, \ldots, X_k$ with their observed values $\boldsymbol{x} = x_1, \ldots, x_k$, and regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$, then the associated MLR models can be obtained and, since $P(Y = \text{VN}|\boldsymbol{x}) + P(Y = \text{N}|\boldsymbol{x}) + P(Y = \text{Neu}|\boldsymbol{x}) + P(Y = \text{P}|\boldsymbol{x}) + P(Y = \text{VP}|\boldsymbol{x}) = 1$, we derive probabilities as

$$P(Y = \text{VN}|\boldsymbol{x}) = \frac{\exp(y_1)}{1 + \exp(y_1) + \exp(y_2) + \exp(y_3) + \exp(y_4)}, \quad (1)$$

$$P(Y = \text{N}|\boldsymbol{x}) = \frac{\exp(y_2)}{1 + \exp(y_1) + \exp(y_2) + \exp(y_3) + \exp(y_4)}, \quad (2)$$

$$P(Y = \text{Neu}|\boldsymbol{x}) = \frac{1}{1 + \exp(y_1) + \exp(y_2) + \exp(y_3) + \exp(y_4)}, \quad (3)$$

$$P(Y = \text{P}|\boldsymbol{x}) = \frac{\exp(y_3)}{1 + \exp(y_1) + \exp(y_2) + \exp(y_3) + \exp(y_4)}, \quad (4)$$

$$P(Y = \text{VP}|\boldsymbol{x}) = \frac{\exp(y_4)}{1 + \exp(y_1) + \exp(y_2) + \exp(y_3) + \exp(y_4)}, \quad (5)$$

where $y_1 = \log(P(Y = \text{VN}|\boldsymbol{x})/P(Y = \text{Neu}|\boldsymbol{x}))$, $y_2 = \log(P(Y = \text{N}|\boldsymbol{x})/P(Y = \text{Neu}|\boldsymbol{x}))$, $y_3 = \log(P(Y = \text{P}|\boldsymbol{x})/P(Y = \text{Neu}|\boldsymbol{x}))$, and $y_4 = \log(P(Y = \text{VP}|\boldsymbol{x})/P(Y = \text{Neu}|\boldsymbol{x}))$, with $y_j = \beta_{0,j} + \beta_{1,j}x_1 + \cdots + \beta_{k,j}x_k$ and $j \in \{1, 2, 3, 4\}$. The equations expressed in (1), (2), (3), (4), and (5) form the foundation for modeling the sentiment classes in our MLR framework. Following the derivation of the five-class sentiment models, we proceed with its fine-tuning and assessment using maximum likelihood estimation for fitting the MLR model.

The fitted model is evaluated using the likelihood ratio test, which compares two models with the ratio of their likelihood functions—one for the reduced model (with only an intercept) and another for the full model (with all covariates) [38]. Evaluating the model fit, pseudo-$R^2$ metrics delineated in [39–41] are used, whose interpretation can be ambiguous as indicated in [42,43]. However, these metrics are important when comparing models, with a model whose superior pseudo-$R^2$ being judged as optimal fit, positing that a pseudo-$R^2$ between 0.2 and 0.4 is a good model fit [44]. Significance of the model coefficients is examined with the Wald test, while the ROC-AUC evaluates the model discriminative capacity. The ROC-AUC ranges from zero to one, where a ROC-AUC of 0.5 suggests the model lacks discriminative ability. For $0.5 <$ ROC-AUC $< 0.7$, the discriminative power is weak; for $0.7 \leq$ ROC-AUC $< 0.8$, it is acceptable; for $0.8 \leq$ ROC-AUC $< 0.9$, it is good; and, for ROC-AUC $\geq 0.9$, the discriminative power is exceptional. Note that, unlike other statistical procedures, MLR requires careful attention to the sample size, especially in the presence of potential collinearity among the covariates. Care is needed with small sample sizes and highly correlated covariates, as these can lead to incorrect or unreliable inferences based on the fitted regression model [45]. As a general guideline, for maximum likelihood estimation in LR models, it is recommended to consider at least 100 cases [46].

### 2.4. Ordinal Regression Models

OR models are a type of regression analysis used when the dependent variable, or target variable, is ordinal with multiple ordered classes. Unlike nominal classes, ordinal classes have a specific order (say, for example, VN, N, Neu, P, VP), but the distance between the classes is not known [37,47]. Before fitting OR models, we evaluate the association between the response variable and covariates using the Cochran–Mantel–Haenszel (row mean scores) statistic. This statistic investigates the relationship between the ordinal response variable and a specific covariate, while adjusting for the effect of another covariate, which is considered as a stratification variable. The ordinality of the response variable is incorporated by assigning scores to response classes, computing means, and inspecting location shifts of means across row levels or sub-populations (formed when covariate levels are cross-classified). Moreover, to account for ordinality in the face of uncertainly equally spaced y-target classes, we assign modified ridit scores [48].

### 2.5. Proportional Odds Model

The proportional odds model, also known as the cumulative logit model, is helpful for grouped continuous variables. Considering a $c$-point scale, where the response classes are $y_1, \ldots, y_c$ and the observed covariates are $\boldsymbol{x} = (x_1, \ldots, x_k)$, we form cumulative probabilities from $P(Y = y_j) = p_j$, for $j \in \{1, \ldots, c\}$. The probability of a response in class $y_j$ or below is given by $P(Y \leq y_j) = p_1 + \cdots + p_j$. This cumulative probability increases with the class, that is, $P(Y \leq y_1) \leq \cdots \leq P(Y \leq y_c) = 1$, reflecting the ordering in the classes. The proportional odds model relates these cumulative probabilities and the covariates using

$$\text{logit}\big(P(Y \leq y_j)\big) = \alpha_j - \boldsymbol{\beta x}, \tag{6}$$

where $\text{logit}(p) = \log(p/(1-p))$ is the logit function, $P(Y \leq y_j)$ the cumulative probability of the target variable being in class $j$ or lower, $\alpha_j$ the threshold parameter for class $j$, and $\boldsymbol{\beta}$ a vector of coefficients for the observed features $\boldsymbol{x}$. The minus sign in front of $\boldsymbol{\beta x}$ in the expression formulated in (6) indicates that, as the covariates increase, the cumulative probability decreases, maintaining the ordering of the classes. The aspect of the proportional odds model is captured by the constant odds ratio for any two sets of observed covariates $\boldsymbol{x}$ and $\boldsymbol{x}'$ presented as (where $\text{log-odds}(p) = \text{logit}(p)$)

$$\frac{\text{odds}\big(P(Y \leq y_j | \boldsymbol{x})\big)}{\text{odds}\big(P(Y \leq y_j | \boldsymbol{x}')\big)} = \frac{\exp(\boldsymbol{\beta x})}{\exp(\boldsymbol{\beta x'})} = \exp(\boldsymbol{\beta}'(\boldsymbol{x} - \boldsymbol{x}')), \tag{7}$$

The constancy stated in (7) applies across all classes $j$, meaning a unit covariate increase universally impacts the odds of a lower class. It is important to validate this assumption when using the model.

### 2.6. Multinomial Logistic versus Ordinal Regression

If the classes have a clear order, OR might be more suitable since it takes this order into account. In addition, MLR treats each class as a separate class without considering any order or hierarchy. Hence, OR might capture the nature of SA better than MLR. The OR model can give meaningful results about shifts from one class to another, considering the order. In comparison, MLR treats shifts from VN to N the same as from VN to VP. This can be less intuitive in a SA context, where we typically think of sentiment as a continuum from negative to positive. The performance of both models can depend on the specifics of the data. MLR might perform better when the assumptions of OR do not hold, or when the order of classes is not very meaningful for predicting the outcome. However, when the ordinal nature of the classes is important, OR can outperform MLR. OR requires the assumption of proportional odds, meaning that the odds are the same regardless of the cut-off point chosen to distinguish between lower and higher ratings. Checking and handling violations of this assumption can add complexity when using OR. In contrast, MLR does not make this assumption, which can make it easier to employ.

### 2.7. Support Vector Machines

SVMs are a supervised ML algorithm used for classification and regression [49,50]. In the context of SA, SVMs are effective due to their capability of handling high-dimensional and sparse data, which is typical in text classification. SVMs perform well with high-dimensional data, making them suitable for text data with large vocabularies. They work effectively with sparse data and have shown superior performance in such situations. SVMs can handle multi-class problems using strategies like one-versus-one and one-versus-all. They are robust against overfitting, especially in high-dimensional spaces. The primary idea of SVMs is to find a hyperplane that best separates the data into classes. For linearly separable data, SVMs find the line that maximizes the margin between classes. For non-linearly separable data, SVMs use the kernel trick to transform data into a higher-dimensional space where it can be separated by a hyperplane. In practice, SVMs introduce the concept of a soft margin to handle noisy data, allowing some misclassifications controlled by a regularization parameter. SVMs are also used for regression tasks, where the goal is to fit as many instances as possible within a margin while minimizing the instances outside the margin, controlled by a parameter.

### 2.8. Random Forests

RF is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees [51,52]. The RF algorithm involves randomly selecting samples with replacement bootstrap samples from the dataset. For each bootstrap sample, a decision tree is grown randomly selecting features out of all features and using the best split on these features to split the node, maximizing reduction in variance (for regression) or Gini impurity (for classification). For prediction, the new object is passed through all trees, and the most voted class is chosen as the prediction. A crucial hyperparameter of RF is the number of trees. More trees reduce overfitting but increase computational complexity. RFs are robust to outliers, scalable, and handle high-dimensional datasets well, which is helpful in SA when using feature vectors like bag-of-words. However, RFs can be less interpretable than models like LR and are more intensive computationally. They may not perform as well as linear models or SVMs when dealing with sparse text data.

### 2.9. An Overview of XGBoost

XGBoost [53] is as an advanced embodiment of the gradient tree boosting algorithm, which integrates regularization to prevent overfitting. Distinct from other boosting algorithms, XGBoost consolidates the contributions of weak learners—specifically, decision trees—to construct a robust predictive model. These trees optimize a loss function, and each successive tree addresses the residuals not captured by its predecessors. A characteristic of XGBoost is the use of regularization, which curbs the model propensity to fit excessively to the training data and may include noise and outliers, augmenting its generalization to unobserved data. Regularization is incorporated within the loss function, directing the model towards simpler configurations and diminishing risk of overfitting. The XGBoost operates as a meta-classifier, amalgamating weak learners into a formidable collective learner. Given a training dataset $(\boldsymbol{x}_i, Y_i)$, where $\boldsymbol{x}_i = x_{i1}, \ldots, x_{ik}$ are the covariates and $Y_i$ the response for the $i$-example, with $i \in \{1, \ldots, n\}$, XGBoost harnesses a set of individual classifiers to forecast the outcome $\hat{Y}_i$. The predictive relationship is expressed as $\hat{Y}_i = \sum_{j=1}^{m} f_j(\boldsymbol{x}_i)$, where $m$ is the number of decision trees and $f_j$ denotes the $j$-th decision tree, with the scores assigned to its leaves. The score for each decision tree is computed through

$$L(j) = \sum_{i=1}^{n} l\left(Y_i, \hat{Y}_i^{(j-1)} + f_j(\boldsymbol{x}_i)\right) + \Omega(f_j), \tag{8}$$

where $l$ is the loss function, $\hat{Y}_i^{(j-1)}$ the prediction for $\boldsymbol{x}_i$ at the $(j-1)$-th iteration, and $\Omega$ the regularization term which constrains the magnitude of the scores on the leaves.

The function $f_j(\boldsymbol{x}_i)$ that minimizes the equation given in (8) is integrated into the decision tree ensemble, culminating in the final classification model. The efficacy of the XGBoost algorithm is attributable to its computation speed and performance metrics. Its elevated performance is due to the system design and adoption of a column block structure, facilitating parallelization during the model construction. Moreover, its capability of managing sparse data and accommodating missing values enhances its versatility and robustness. In summary, the XGBoost classifier harnesses the principles of gradient boosting while incorporating regularization to achieve high efficiency in various complex predictive modeling tasks. This robust learning apparatus, with its strong theoretical underpinnings and empirical validation, continues to serve as a pivotal tool in the ML arena.

### *2.10. Comparison between Random Forests and XGBoost*

Both XGBoost and RF are ensemble ML algorithms that utilize decision trees as their base learners. However, the methodologies underlying their learning processes, approaches to the bias–variance tradeoff, and strategies for evaluating feature importance display fundamental differences. RFs construct each tree independently, leveraging the averaging of results to reduce variance and mitigate the risk of overfitting. In contrast, XGBoost builds trees sequentially, with each successive tree amending the errors of its predecessor. This progressive refinement inherent in boosting algorithms can heighten sensitivity to noise and outliers, but also has the potential to improve model accuracy with meticulous tuning.

While RFs inherently manage the variance by amalgamating independently constructed decision trees, XGBoost addresses both bias and variance through the creation of a robust predictive model via a stepwise, additive, and sequential process. Provided that the hyperparameters are finely tuned, this typically confers upon XGBoost a higher level of accuracy compared to RFs. In the context of feature selection, RFs integrate a random subset of features when bifurcating nodes during tree construction, fostering model diversity and further reducing overfitting. Conversely, XGBoost contemplates all available features for node splits, exerting control over model complexity through regularization parameters. The two algorithms also differ in their computation of feature importance. In RFs, importance is deduced from the aggregate decrease in node impurities resulting from splits on the feature, averaged across all trees. XGBoost, however, ascertains feature importance based on the average gain of the feature when utilized in trees. Considering raw performance metrics, XGBoost frequently surpasses RF. Owing to its gradient boosting framework, XGBoost is often capable of delivering superior predictive accuracy. Nevertheless, the sequential training process inherent in XGBoost may necessitate a more meticulous tuning phase compared to that required for RFs.

### *2.11. Performance Metrics and Validation Techniques*

Evaluating the predictive performance and reliability of a trained model is critical to ensure its practical applicability. This subsection outlines the assessment framework, which includes a suite of metrics and validation methodologies, with a particular focus on cross-validation and the train–test split paradigm, as detailed in [4]. Precision and recall (also known as sensitivity) are crucial metrics for assessing the effectiveness of model predictions for the positive class. Precision is defined by the ratio of true positive (TP) instances to the sum of TP and false positive (FP) instances, which can be quantitatively expressed as $\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$. Conversely, sensitivity is defined using TP and false negatives (FN) as $\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$. Accuracy, another cornerstone metric, quantifies the overall proportion of accurate predictions made by the model, computed as $\text{Accuracy} = \text{True predictions}/\text{Total predictions}$. For binary classification tasks, the accuracy metric is specified as $\text{Accuracy (binary)} = (\text{TN} + \text{TP})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$, where TN is true negatives. These metrics are coalesced within the confusion matrix, which is presented in Table 1.

**Table 1.** Confusion matrix scheme.

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| Observed Positive | TP | FN |
| Observed Negative | FP | TN |

The F1-score approaches to unity exclusively when both precision and sensitivity are perfect. F1-score only attains high values when both metrics are concurrently high, so providing a more nuanced performance indicator than accuracy alone. To assess the model stability and predictive consistency, a resampling procedure named *k*-fold cross-validation is employed. This procedure partitions the dataset into *k* equally sized segments, consecutively utilizing each segment as a validation set while the remainder function as the training set.

Complementing these metrics, learning curves offer vital insights into a model learning trajectory and data utilization efficiency. By graphically charting a risk or cost function against the volume of training data, learning curves facilitate the determination of the data quantity requisite for optimizing model performance.

## 3. Dataset and Its Preparation for Analytics

In this section, we detail the dataset, the data cleansing process, and the methodology used.

### *3.1. Dataset Description*

This study employs a dataset encompassing customer support interactions, systematically gathered since 2021. Our analysis focuses on a one-year subset of messages, including data fields such as message content, message number, type, and associated sentiment classifications. Message lengths diverge significantly, ranging from concise 20-word expressions to extensive reviews exceeding 200 words.

The original dataset comprised 58,172 records. Data cleansing resulted in the exclusion of 524 duplicate entries, 11,996 unclassified records, and 654 records with identical messages. Consequently, the sanitized dataset consisted of 44,998 unique cases. It is important to specify that SA on this dataset was pre-classified by a consultancy previously engaged by the company.

While the specific criteria utilized by the consultancy remain proprietary and confidential, the reliability of the classifications is endorsed by the satisfaction of our business client with the analytical outcomes. This endorsement serves as a testament to the validity and business relevance of the pre-classified sentiments. The sentiments were categorized into five distinct classes, varying from very negative to very positive. In the cleansed dataset of approximately 45,000 cases, 78% are designated as service requests (SRs), 14% as information requests (IRs), 8% as complaints, and the remaining entries are dispersed among compliments and suggestions. Figure 1 illustrates the distribution of these classes.

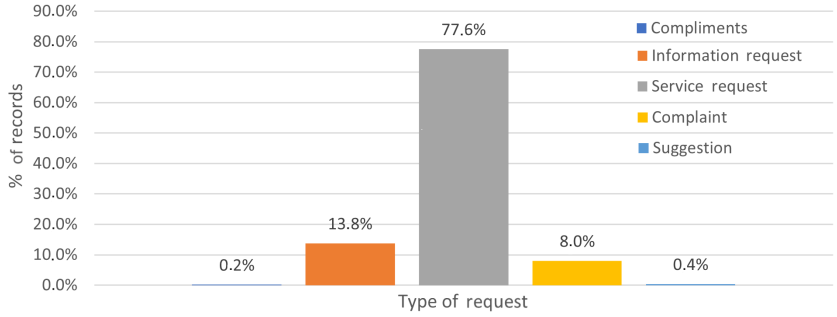



**Figure 1.** Distribution of request types in the Portuguese data.

Regarding sentiment distribution, 79% of the entries are neutral. This proportion aligns with the predominance of SRs and IRs, which typically elicit a neutral response. Additionally, 20% of the records manifest negative sentiment, whereas the balance consists of very negative, positive, and very positive sentiments. Notably, records with very positive sentiments are scarce, suggesting customer reticence in expressing positive feedback. Figure 2 delineates the sentiment distribution.

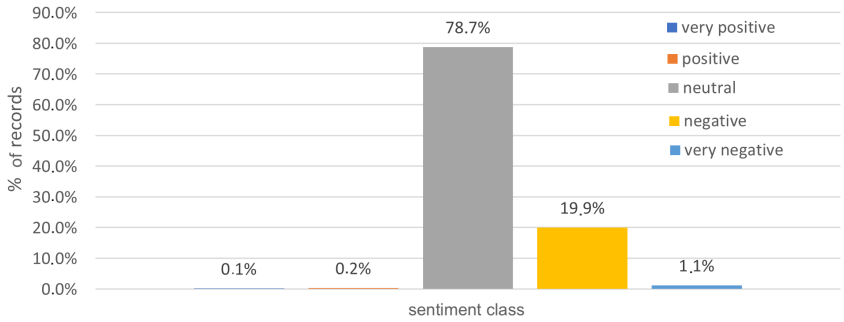



**Figure 2.** Distribution of sentiment classes in the Portuguese data.

Table 2 reveals a strong correlation between sentiment classes and request types. Positive sentiments are chiefly associated with compliments, SRs, and IRs, while negative sentiments correlate more with complaints. It should be noted that approximately 10% of records classified as positive are actually complaints, indicating an incongruence. Further scrutiny suggests that this incongruence may stem from the deployment of irony or the incorporation of ostensibly positive lexemes from Portuguese like "bom dia" (good morning), "agradeço" (thank you), and "cumprimentos" (regards), in otherwise negative messages.

The marginal totals in Table 2 represent observed counts of records classified under each sentiment class, not percentages. These counts are pivotal in understanding the distribution of sentiment classifications across different request types without normalization by the total number of records. Notably, the marginal total for the "Very Positive" class is 24, indicating the observed number of records classified as such. It is important to recognize that some "Very Positive" sentiments may be associated with unique types of interactions that do not fit neatly into the conventional classes of SR, IR, complaints, or suggestions. This can lead to less visible representations within these predefined classes and explains why some data points might not align perfectly with the standard classifications. Such distinctions underscore the complex nature of SA and highlight the need to consider all nuances in our evaluation.

**Table 2.** Percentage of request types by sentiment class with record counts in the final column of Portuguese data.

| | Compliment | IR | SR | Complaint | Suggestion | Record Count |
|---|---|---|---|---|---|---|
| Very Positive | 33.3% | 12.5% | 20.8% | 0.0% | 0.0% | 24 |
| Positive | 9.8% | 29.4% | 50.0% | 9.8% | 1.0% | 102 |
| Neutral | 0.2% | 14.9% | 78.4% | 6.1% | 0.4% | 35,412 |
| Negative | 0.0% | 9.2% | 75.5% | 14.8% | 0.5% | 8964 |
| Very Negative | 0.0% | 11.5% | 64.5% | 23.4% | 0.6% | 496 |

*3.2. Data Cleaning Procedure*

Data cleaning is a crucial step in preparing the text messages for SA. Inaccurate or unclean data can greatly distort the results and decrease the validity of the SA. Therefore, a rigorous cleaning process is employed. Consider the following raw text message received from a customer, which is used to illustrate the cleaning procedure:

> "Bom dia,\ngostaria de saber quando terão cadeiras, novamente, disponiveis.\nMuito obrigada!\n https://www.continente.pt/produto/cadeira-echair-confort-kasa-284190.html\n\nMelhores cumprimentos,\n\nMaria Silva, mariasilva@gmail.com \n\nDe: Continente ajuda@continente.pt\nEnviado: 25 de setembro de 2021 12:06 \nPara: Maria Silva mariasilva@gmail.com\nAssunto: Confirmação de Encomenda\n\n"

Translation to English:

> "Good morning,\nI would like to know when you will have chairs available again.\nThank you very much!\n https://www.continente.pt/produto/ cadeira-echair-confort-kasa-284190.html\n Best regards,\nMaria Silva, mariasilva@gmail.com \n\nFrom: Continente ajuda@continente.pt\nSent: September 25, 2021 12:06 \n\nTo: Maria Silva mariasilva@gmail.com\nSubject: Order Confirmation\n\n"

↓

Step 1: Space restoration

Upon import, spaces are automatically replaced with \n. This step restores the original spacing structure for readability as indicated below:

> "Bom dia, gostaria de saber quando terão estas cadeiras, novamente, disponiveis. Muito obrigada! https://www.continente.pt/produto/cadeira-echair-confort-kasa-284190.html Melhores cumprimentos, Maria Silva, mariasilva@gmail.com —- De: Continente <ajuda@continente.pt> Enviado: 25 de setembro de 2021 12:06 Para: Maria Silva <mariasilva@gmail.com> Assunto: Confirmação de Encomenda"

Translation to English:

> "Good morning, I would like to know when you will have these chairs available again. Thank you very much! https://www.continente.pt/produto/cadeira-echair-confort- kasa-284190.html Best regards, Maria Silva, mariasilva@gmail.com —- From: Continente <ajuda@continente.pt> Sent: September 25, 2021 12:06 To: Maria Silva <mariasilva@gmail.com> Subject: Order Confirmation"

↓

Step 2: Character removal and case standardization

All characters except for '.', '!', and '@' are eliminated. Case standardization is achieved by converting all characters to lowercase as indicated below:

> "bom dia gostaria de saber quando terão estas cadeiras novamente disponiveis. muito obrigada! melhores cumprimentos Maria Silva mariasilva@gmail.com de continente enviado de setembro de para maria silva assunto confirmação de encomenda"

Translation to English:

> "good morning i would like to know when you will have these chairs available again. thank you very much! best regards maria silva mariasilva@gmail.com from continente sent september to maria silva subject order confirmation"

↓

Step 3: Retailer message exclusion

Messages from the retailer are excluded to concentrate solely on the customer sentiment as indicated below:

> "bom dia gostaria de saber quando terão estas cadeiras novamente disponiveis. muito obrigada! melhores cumprimentos maria silva mariasilva@gmail.com"

Translation to English:

> "good morning i would like to know when you will have these chairs available again. thank you very much! best regards maria silva mariasilva@gmail.com"

↓

Step 4: Email address removal

Email addresses are removed to reduce noise and distractions as indicated below:

> "bom dia gostaria de saber quando terão estas cadeiras novamente disponiveis. muito obrigada! melhores cumprimentos maria silva"

Translation to English:

> "good morning i would like to know when you will have these chairs available again. thank you very much! best regards maria silva"

↓

Step 5: Accent and period removal

All accented characters and periods are removed for text consistency as indicated below:

> "bom dia gostaria de saber quando terao estas cadeiras novamente disponiveis muito obrigada! melhores cumprimentos maria silva"

Translation to English:

> "good morning i would like to know when you will have these chairs available again thank you very much! best regards maria silva"

↓

Step 6: Greeting elimination

Common greetings are excluded as they do not contribute to sentiment identification as indicated below:

> "gostaria de saber quando terao estas cadeiras novamente disponiveis muito obrigada! melhores cumprimentos maria silva"

Translation to English:

> "i would like to know when you will have these chairs available again thank you very much! best regards maria silva"

↓

Step 7: Stop words removal

Commonly occurring stop words, except for 'não', are removed to focus on more meaningful text elements as indicated below:

"gostaria saber terao cadeiras novamente disponiveis obrigada! melhores cumprimentos maria silva"

Translation to English:

"would like know will have chairs available again thank you! best regards maria silva"

↓

Step 8: Stemming

The stemming process consolidates various forms of a word into a single representative base form as indicated below:

"gost sab ter cade nov disponi obrigada! melhor cumpr mar silv"

Translation to English:

"like know have chair avail thank you! best regards mar silv"

After stemming, the data cleaning process may include one more conditional step. Some messages are received through external platforms such as the Portal da Queixa (Complaint Website). These messages often include additional automated notes which are not relevant to the SA. Therefore, as a final step, these extraneous notes are removed as illustrated below:

"O valor da encomenda nº 000000000 (5 July 2021) foi debitado duas vezes no meu cartão crédito. Fiz o pagamento através do MB Way. Peço a resolução desta situação com a maior brevidade.

---

Nova reclamação recebida
Olá, o Portal da Queixa rececionou uma reclamação por parte de um utilizador dirigida à marca Continente e que contém dados pessoais do(a) reclamante que recolhemos nos termos da nossa Política de Privacidade e Proteção de Dados
portaldaqueixa.com

---

Continente - Valor de encomenda online debitado duas vezes
Reclamação #00000000 em 2021-07-07 18:44:52"

Translation to English:

"The amount of order nº 000000000 (5 July 2021) was charged twice on my credit card. I made the payment via MB Way. I ask for the resolution of this situation as soon as possible.

---

New complaint received
Hello, our website received a complaint from a user addressed to the brand Continente containing personal data of the complainant collected under our Privacy and Data Protection Policy
portaldaqueixa.com

---

Continente - Online order amount charged twice
Complaint #00000000 on 2021-07-07 18:44:52"

↓

"valor encomenda debitado duas vezes cartao credito fiz pagamento atraves mb way peco resolucao desta situacao maior brevidade"

Translation to English:

> "order amount charged twice credit card made payment via mb way ask resolution this situation as soon as possible"

*3.3. Feature Extraction*

Feature extraction serves as a crucial step following the data cleaning process. This step transforms the cleaned text messages into a structured format, allowing for an effective SA. Through the extraction of key features, the data are compressed into a representative form, facilitating the discovery of underlying patterns and relationships. The first group of features targets aspects such as the textual complexity and length of the messages. These aspects can reveal information about the sender engagement level and the message emotional intensity. Next, we describe these features.

Feature 1: Number of words (NW)

This feature captures the message length and can correlate with the intensity of expressed sentiment. Longer messages may contain more emotional cues or elaborate arguments, so influencing the overall sentiment.

Feature 2: Number of letters (NL)

Similar to NW, NL gives insights into the emotional intensity conveyed by the message. Longer messages could indicate more complex emotions, while shorter messages could represent concise responses.

Feature 3: Number of vowels (NV)

Counting the vowels in each message provides an understanding of vocal intensity, which may reflect the emotional emphasis of the content.

Feature 4: Lexical diversity (LD)

This metric represents the proportion of unique words, indicating the lexical richness and possibly the complexity of the expressed sentiment.

Now, we state sentiment lexicon-based features. The SentiLex-PT dictionary, tailored for Portuguese SA, was employed to extract lexicon-based features.

Feature 5: Sentiment score (SS)

This feature is derived from the difference between counts of positive and negative words in the message.

Feature 6: Positive score (PS) and negative score (NS)

This feature captures the frequency of positively and negatively charged words, respectively.

Feature 7: Number of adjectives (NA)

This feature gauges the number of adjectives used, offering insights into the linguistic characteristics and emotional tone of the message.

The integration of these lexicon-based features enhances the robustness and interpretability of our SA, contributing to a comprehensive view of the customer sentiment. The extracted features are vital for characterizing and representing the cleaned data, so laying the foundation for a nuanced SA. The combination of text complexity and lexicon-based features allows for a multidimensional view of the data, which is instrumental for effective SA.

### 3.4. Text to Number Conversion

In most ML tasks, including SA, numerical data form the core input for algorithms. To adapt the text-based customer messages for computational analysis, it is indispensable to transform them into a numerical representation. The conversion was performed using `Python` 3.8, leveraging the `Tokenizer` class available in the `Keras` library, version 2.4.0. The code was executed on a system with an Intel Core i7 processor and 16 GB of RAM, ensuring consistent and efficient computational performance. The `Tokenizer` class dissects the textual input into individual tokens (words, in this context) and maps each unique token to a distinct integer. This results in a binary vector for each customer message, where each entry indicates the presence (1) or absence (0) of a particular word in the text.

Example:

Consider the sentence "gosto de ir ao continente" (I like to go to Continente). After tokenization and vectorization, this sentence is represented as a binary vector expressed in Table 3. In the converted dataset, each customer text message is replaced by its corresponding binary vector, enabling the ML model to effectively perform SA.

**Table 3.** Binary vector representation of the sentence "gosto de ir ao Continente" (I like to go to Continente) such as it appears in the Python data frame.

| **Continente** | **Entrega (Delivery)** | **ir (go)** | **muito (very)** | ... | **de** | **gosto (I like)** | **ao (to)** |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | ... | 1 | 1 | 1 |

### 3.5. Techniques for Addressing Class Imbalance

The problem of class imbalance is a well-acknowledged issue in ML disciplines. This asymmetric distribution of classes hampers both model training and evaluation. For the dataset under consideration, a disproportionate distribution among positive, neutral, and other classes necessitates strategies for balance correction. The selection of a class balancing method is empirical, depending on model architecture, data distribution, and the specific challenges of the learning task. Various methods are discussed next.

Oversampling:

This increases the size of the minority class by replicating instances or generating synthetic samples and enhancing the model performance for the minority class.

Undersampling:

This reduces the size of the majority class either through random deletion or advanced methods like Tomek links, resulting in a more balanced class distribution.

Hybrid method:

This combines oversampling and undersampling to achieve equilibrium in class distribution.

Class weight adjustment:

This alters class weights during training to emphasize the minority class in the loss function, reinforcing its importance in the learning process.

### 3.6. Utilization of N-gram Analysis for Sentiment Decoding in Retail

*N*-gram analysis emerges as a pivotal technique in the domain of SA, particularly for retail enterprises seeking to decode the intricate fabric of customer sentiment expressed in feedback. An *N*-gram refers to a contiguous string of *N* words extracted from a given text, where *N* denotes the cardinality of the word sequence. The application of *N*-gram analysis transcends beyond mere frequency counts of individual words, enabling the discovery of syntactical and semantical relationships amongst words. This aids in discerning more complex and context-rich expressions, thereby unveiling intricate patterns and associations germane to customer sentiment.

The employment of *N*-gram analysis serves a dual purpose: (i) it helps in identifying generic lexemes indicative of sentiment and (ii) it also identifies lexico-syntactic patterns idiosyncratic to the retail sector. Such linguistic formulations are often laden with insights into customer predilections, bottlenecks, and shifting emotional tides over temporal gradients. By meticulously analyzing the sentiment polarity affiliated with specific *N*-grams, retail businesses can gain a more nuanced understanding of customer sentiment landscape. These empirically derived insights can subsequently guide a gamut of strategic decisions, ranging from product innovation to the customization of marketing strategies, thereby ensuring congruence with evolving customer sentiment.

## 4. Empirical Evaluation and Discourse

In this section, we present an analysis of various ML architectures, both custom-trained and pre-configured, for SA in a Portuguese retail customer service.

### *4.1. Methodology Overview*

The selection of models for SA was driven to compare traditional ML approaches with modern NLP-driven algorithms, aiming to identify the most effective tools for accurate and reliable SA in a Portuguese retail customer service. The initial phase of the evaluation encompassed a broad spectrum of models to ensure a comprehensive view of computational approaches. Custom-trained models included NB, RF, MLR, XGBoost, and OR, while pre-configured models comprised `TextBlob`, `Vader`, and `Twitter Roberta`. Models were categorized and selected based on their computational methodologies and suitability to the data characteristics as follows:

- [Traditional ML models] NB, RF, and XGBoost were chosen for their established reliability in text classification tasks.
- [Advanced regression models] MLR and OR were employed to effectively address the ordered nature of sentiment classes.
- [Pre-configured NLP models] `TextBlob`, `Vader`, and `Twitter Roberta` were included to assess the effectiveness of models pre-trained on extensive diverse datasets, primarily in English.

The methodology used in our study is illustrated in Figure 3, which outlines the various stages of our analysis, from data collection to the final compilation of results.

For the preliminary analysis, a feature set constituting 40,155 unique lexemes was employed across all listed models. Initial observations revealed a deficiency in the model predictive capability, which led to data consolidation and the enactment of multiple experimental paradigms.

To further investigate these deficiencies, several pre-processing and resampling techniques were benchmarked, including joining positive classes, oversampling, synthetic minority oversampling technique (SMOTE), undersampling, removing stopwords, and stemming. These techniques were applied in various combinations to identify the optimal pre-processing strategy for all models. Notably, the minority class experienced significant improvements in classification accuracy through random oversampling. Additionally, lexical simplification by removing stopwords further optimized the models, as illustrated in Figure 4. After this preliminary evaluation, the analysis was concentrated on models that demonstrated superior performance, particularly MLR and XGBoost, due to their higher accuracy and interpretability. These models were chosen because they consistently outperformed others in terms of key performance metrics such as accuracy, ROC-AUC, and precision–recall balance. This allowed for a more detailed and impactful analysis, reducing complexity and enhancing the relevance of the findings.

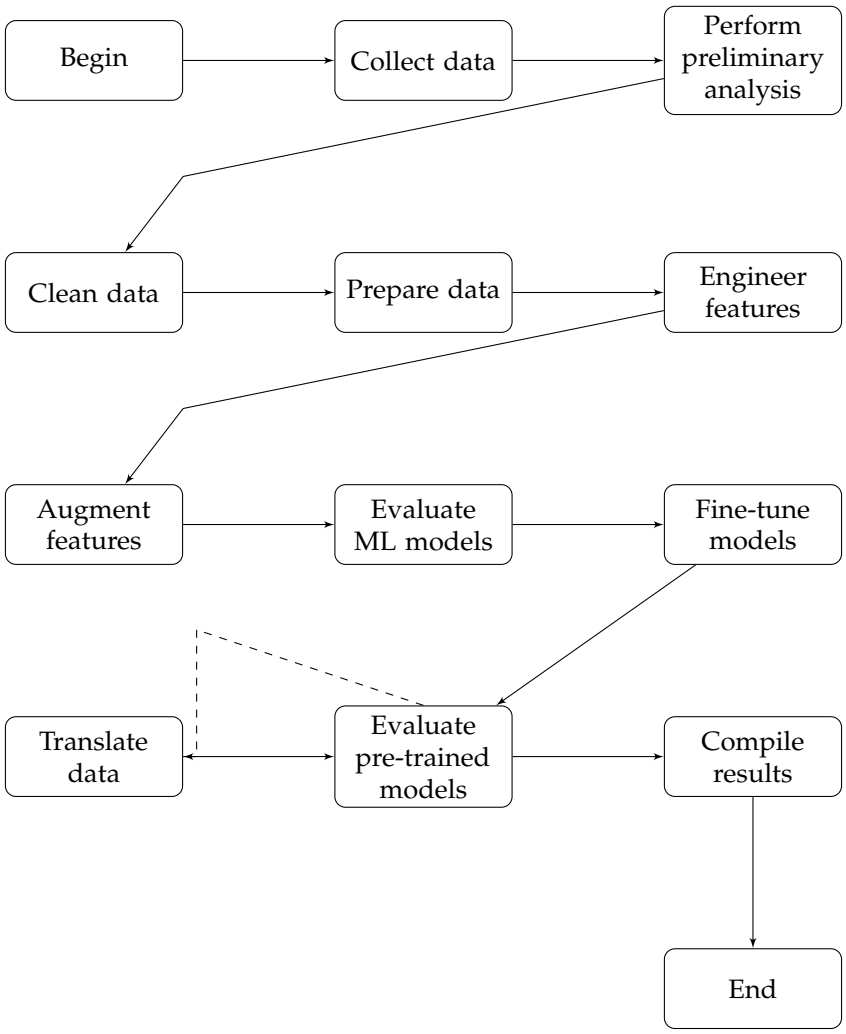



**Figure 3.** Detailed flowchart of the presented methodology.

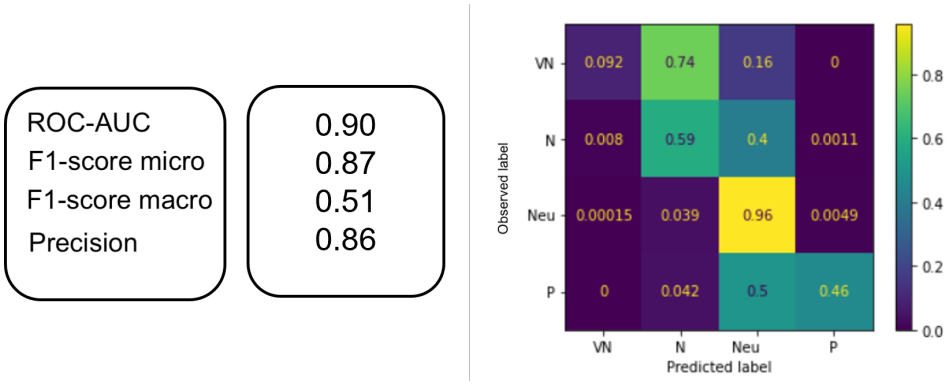



**Figure 4.** Performance of the MLR model with the addition of SS, PS, and NS features in the Portuguese data.

### *4.2. Fine-Tuning and Feature Augmentation*

Following the identification of MLR and XGBoost as the most promising models, we proceeded with subsequent experimental phases to fine-tune these models and improve their performance through feature augmentation. The feature set was augmented by incorporating three additional lexical attributes: SS, PS, and NS, so enhancing the initial feature space of 40,155 unique terms. The incorporation of an additional feature, NA, culminated in a demonstrable uplift in the ROC-AUC score, as substantiated in Figure 5.

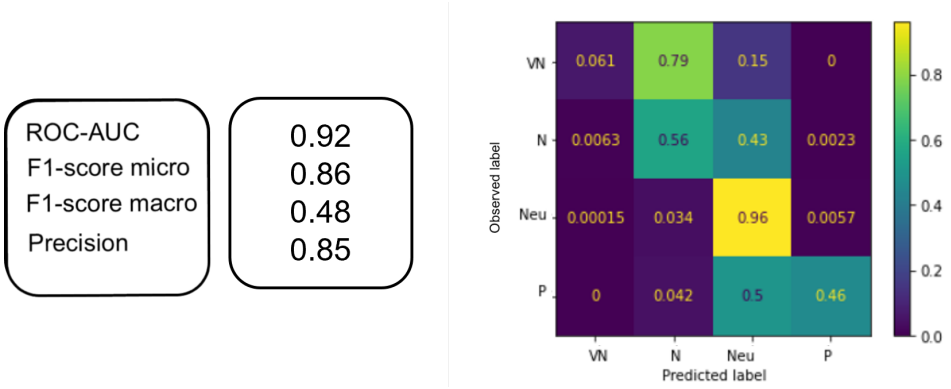



**Figure 5.** Performance of the MLR model after adding the NA feature in the Portuguese data.

During subsequent evaluations, we examined the high dimensionality of the feature space in relation to the dataset size. To evaluate the model robustness against noisy and infrequent terms, we conducted further analyses. These analyses, as shown in Figure 6, demonstrated that the removal of less frequent words led to improved model performance, increasing an 8% in predictive capacity for the positive class. This demonstration shows the importance of feature selection in enhancing model performance.

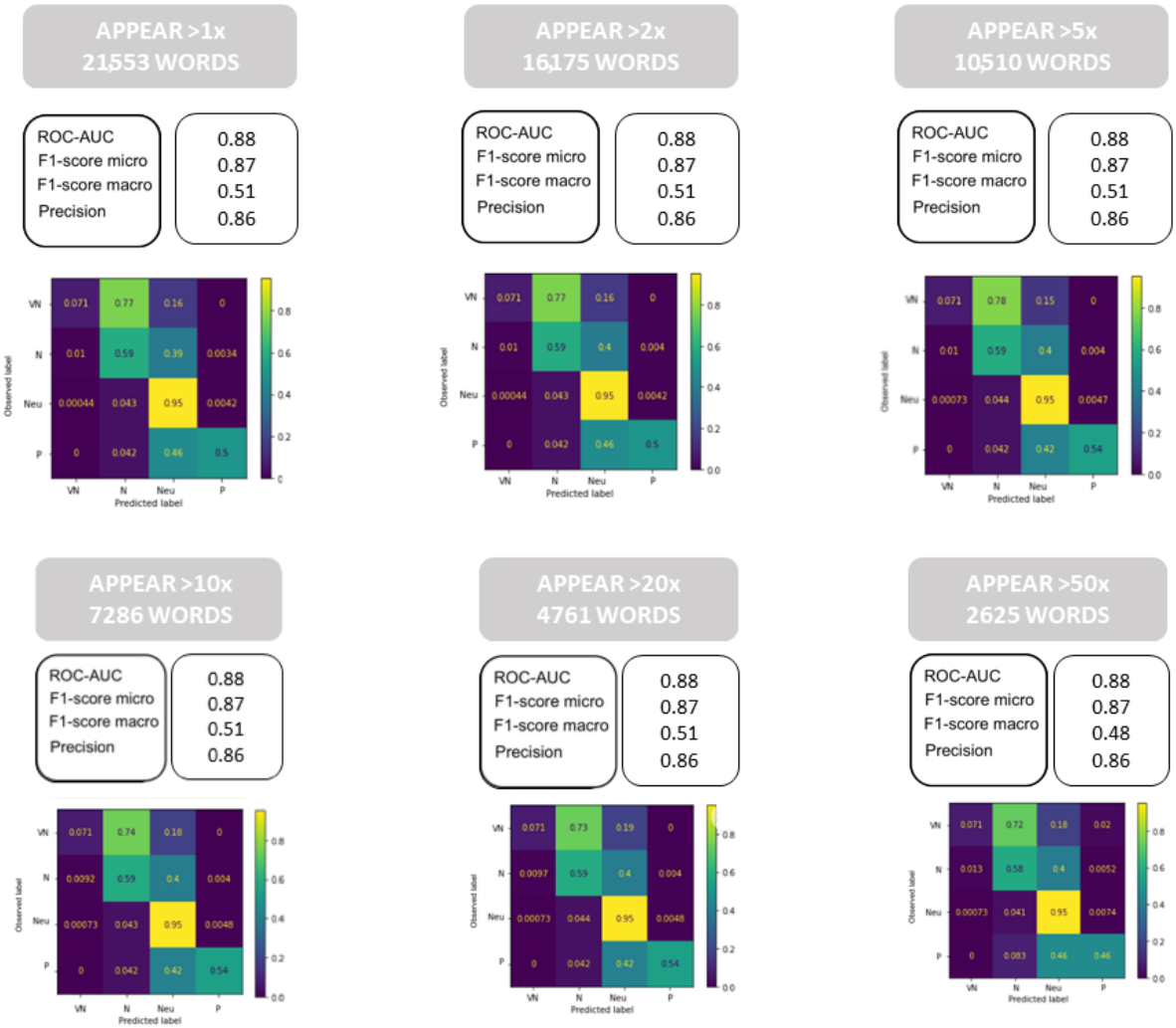



**Figure 6.** Impact on model performance after removal of less frequent words from the features with the Portuguese data.

During the evaluation phase for the MLR model, attention was given to the balanced class parameter. As illustrated in Figure 7, activating this parameter improves the model handling of class imbalances, regardless of the use of oversampling techniques. Despite significant improvements in both the ROC-AUC score and precision metrics, these two metrics were not applied simultaneously due to a slight reduction in predictive capacity for the positive class. Specifically, in the negative class, the model predictive capability increased from 59% to 74%, whereas the very negative class increased from 1% to 22%. Upon concluding the tests on the MLR model, the next step was to assess the impact of additional features, that is, NW, NL, NV, and LD, as shown in Figure 8. We then proceeded to test the XGBoost model. The results of this test are presented in Figure 9. For this purpose, the features that yielded the best results in the MLR model were considered, specifically, 7286 features. A similar test was conducted for the XGBoost model, using the oversampling technique and the balanced class hyperparameter.

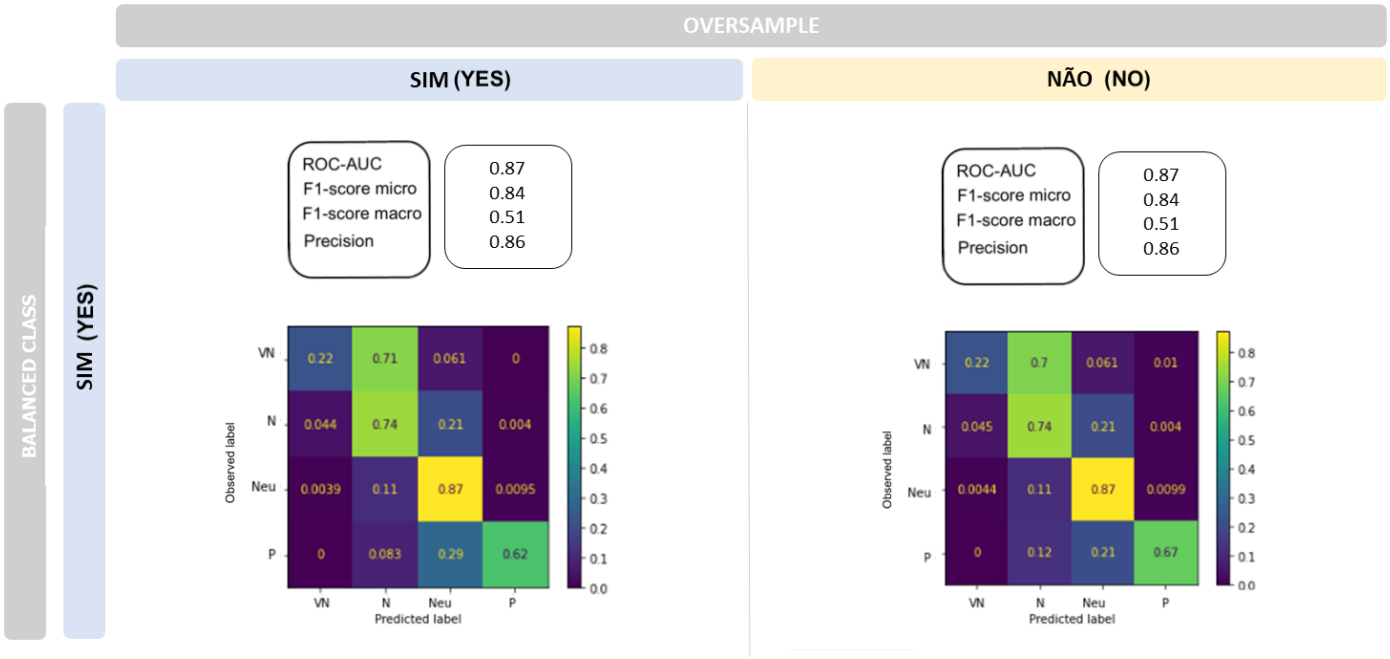



**Figure 7.** Comparison of performance with and without the balanced class parameter with Portuguese data.

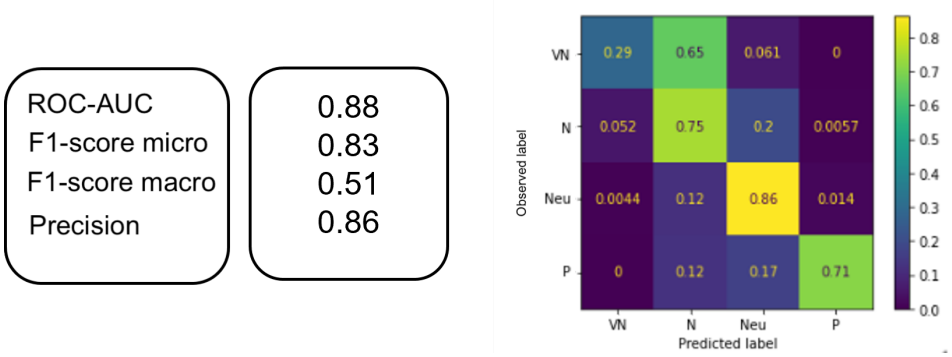



**Figure 8.** Performance of the MLR model with optimized features with the Portuguese data.

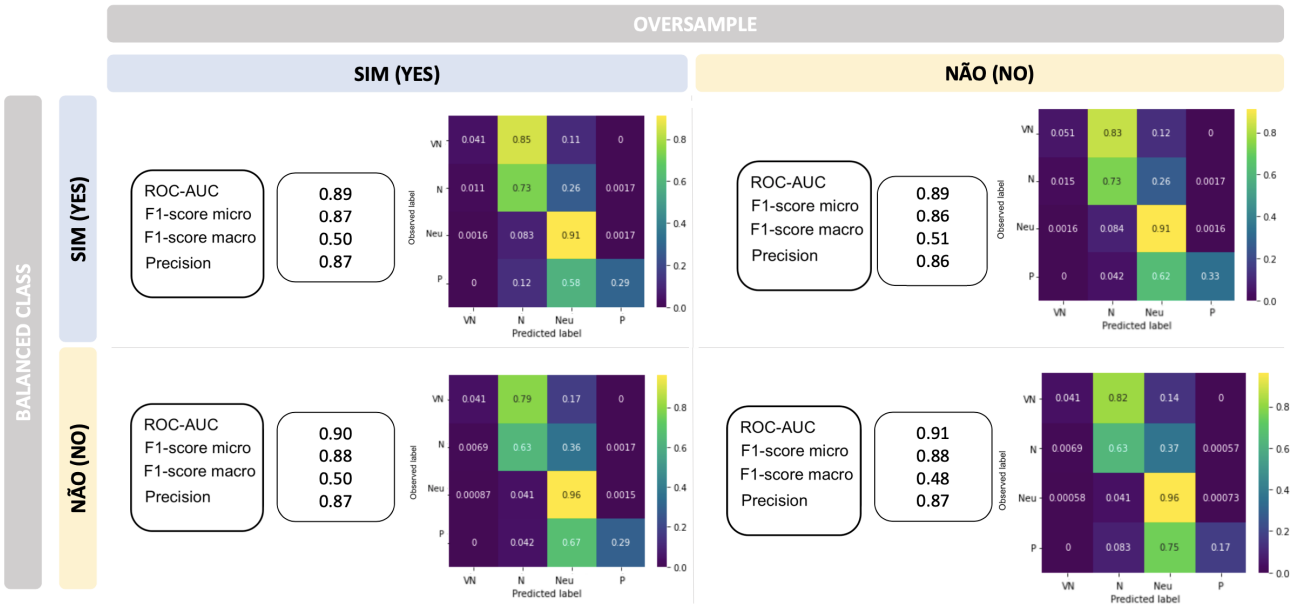



**Figure 9.** Performance of the XGBoost model with optimal features and the Portuguese data.

### *4.3. Evaluating Pre-Trained English Models*

Next, we focus on the application of pre-trained English models for SA. Given that the original data were in Portuguese, accurate translation into English was imperative to ensure the applicability of these models. To accomplish this, we utilized the `Translator` function from the `googletrans` library of `Python`. While this function is generally reliable, we are aware of the potential for semantic loss or distortion during automatic translation. To mitigate this distortion, we implemented several checks:

- [Validation of sample translations] A subset of translations was manually checked to ensure that the translated text retained the original sentiment and meaning.
- [Contextual adjustments] Where necessary, contextual knowledge was applied to adjust translations that might be literally correct but contextually inappropriate, especially in the case of idiomatic expressions.
- [Consistency checks] Consistency across translations was ensured by checking the outputs for commonly recurring phrases and terms, adjusting the translation model parameters when discrepancies were found.

These checks helped in maintaining the integrity of SA when applying English-based SA models to Portuguese text. Due to the large volume of data, the `Translator` function becomes computationally expensive and slow. Thus, the test was conducted on a subset of 20,000 data points. Also, two of these models were designed to work with only three sentiment classes: N, Neu, and P. For the purpose of our analysis, we combined VN with N and VP with P. We initially evaluated several pre-trained models, including `TextBlob` and `Vader`. `TextBlob`, a powerful library of `Python` for text and NLP tasks, outputs a polarity score ranging from zero to one. However, our findings indicated that `TextBlob` performance may not be optimal for this dataset, particularly in the context of Portuguese retail, as it frequently assigned intermediate polarity scores to messages categorized as negative and inappropriately low scores to those deemed positive. The second pre-trained model evaluated was `Vader`, a popular SA tool from the `NLTK` library of `Python`. While `Vader` excelled in classifying neutral sentiments, it struggled with identifying positive and negative sentiments accurately, which does not align well with the needs of retail applications that require acute sensitivity to more polarized sentiments. Next, we present the `Twitter Roberta`-based sentiment model, trained on a large Twitter dataset. Unlike `Vader`, which excels in classifying neutral sentiments but struggles with negative and positive ones, `Twitter Roberta` shows a strong performance in identifying both negative and positive sentiments but has difficulty with neutral cases, as illustrated in Figure 10.

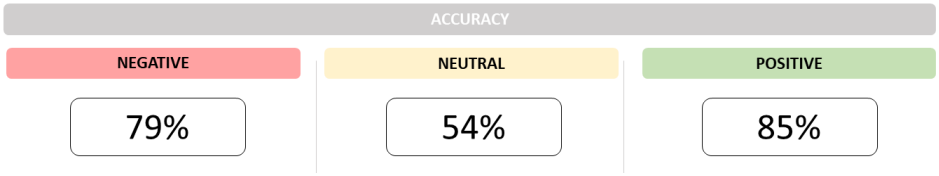



**Figure 10.** Classification performance of the `Twitter Roberta`-based sentiment model with the Portuguese data.

Strong suit of `Twitter Roberta` is its proficiency in classifying negative and positive sentiments. This is particularly advantageous for business applications that require precise sentiment identification to inform strategic decisions, customer engagement, and marketing initiatives. However, the model struggle with neutral sentiments presents a limitation, as misclassification in this area can lead to inaccuracies in SA, potentially skewing business insights. To mitigate this limitation, our methodology incorporates an analysis that effectively differentiates between truly neutral sentiments and those subtly expressing dissatisfaction or mild contentment. This involves a detailed examination of linguistic cues and contextual factors influencing sentiment classification. By enhancing the model sensitivity to these cues, we reduce misclassification and improve the reliability of sentiment assessments, particularly for neutral sentiments, which constitute a significant portion of our data.

Despite its high accuracy for N and P classifications, our analysis suggests that the retailer could benefit from prioritizing feedback that skews slightly negative over typical neutral feedback. Given the high prevalence of Neu cases in the dataset, the `Twitter Roberta`-based sentiment model, along with the traditional MLR model, has proven to be a valuable asset for SA within the customer support team of a Portuguese retailer. `Twitter Roberta` strength in identifying polarized sentiments complements MLR balanced performance and interpretability. This dual-model complement improves the accuracy and reliability of sentiment assessments, supporting more informed and strategic business decisions.

## 5. Conclusions and Discussion

This research undertook the critical task of evaluating a range of machine learning algorithms, both traditional and pre-trained, for their efficacy in the domain of sentiment analysis. Our ultimate objective was to provide a novel and integrated methodology tailored for customer sentiment analysis, which allowed us to identify an optimal computational model that enhanced the customer service operations for a retail business in Portugal.

Initial investigations involved the assessment of traditional algorithms such as naive Bayes, random forests, and multinomial logistic regression. Empirical evidence suggested that multinomial logistic regression excelled in both receiver operating characteristic area under the curve and precision metrics, particularly for positive and neutral sentiment classes. The algorithm also offered reasonable accuracy for negative sentiments.

The effectiveness of the multinomial logistic regression model, particularly in identifying positive and neutral sentiments, provided actionable insights for retail managers. By harnessing these insights, retailers can refine their customer engagement strategies to improve overall customer satisfaction. For example, targeted responses to negative feedback can be structured to mitigate dissatisfaction promptly, whereas reinforcing behaviors that lead to positive feedback may enhance customer loyalty and advocacy.

Advanced feature techniques and data treatments were employed to enhance the model performance further, following the guidelines presented in the extant literature [54,55]. Our methodology emphasized efficient text representation and low-dimensional dense text vectors for capturing nuanced semantic relationships in text. Our analysis also involved strategic feature selection, whereby specific features were either introduced or eliminated based on their performance impact. The implementation of the balanced class parameter notably improved the algorithm capability to handle imbalanced data.

The implications of these findings are particularly significant for the retail sector in Portugal, where understanding nuanced consumer sentiments can directly influence policy and operational strategies. Retail managers are encouraged to use these insights to tailor marketing initiatives and customer service protocols, ensuring they align more closely with the evolving preferences and expectations of Portuguese consumers.

In terms of contemporary machine learning algorithms, the present study expanded to include the extreme gradient boost and ordinal regression models. These models, although robust, failed to surpass the performance levels achieved by the multinomial logistic regression model. Furthermore, the research engaged in the evaluation of pre-trained models, notably `TextBlob`, `Vader`, and `Twitter Roberta` sentiment. Interestingly, the `Twitter Roberta` model displayed superior performance in classifying both negative and positive sentiments, thereby emerging as a potentially invaluable resource for customer service operations. It is crucial, however, to mention the model limitations in identifying neutral sentiments accurately. Such limitations could yield erroneous classifications and so it should be meticulously considered when applying the model in real-world scenarios.

In summary, the present study underscores the applicability of both the multinomial logistic regression model and the `Twitter Roberta`-based sentiment model for sentiment analysis within the customer service framework of a major Portuguese retail business. These models promise high predictive accuracies for classifying a range of sentiments, thereby facilitating more effective customer service operations. It is worth noting that these models have been actively implemented by a leading retailer in Portugal and are continuously refined and adjusted based on ongoing feedback and new data. This practical implementation and iterative enhancement process underscore the model effectiveness and adaptability in real-world settings.

The application of advanced machine learning techniques in sentiment analysis demonstrated the potential for improving customer service and informing policy decisions that impact the retail sector in Portugal. As we continue to refine these models by incorporating broader linguistic and cultural nuances, we anticipate improvements in the predictive accuracy of sentiment analysis tools. This ongoing evolution is expected to extend their applicability across the interaction of diverse market environments, enabling deeper insights into consumer behavior and supporting more nuanced decision-making in both business and policy contexts. The concept of the quadruple helix model is pivotal in understanding these interactions. This model involves collaboration among four key societal stakeholders: academia, industry, government, and civil society, emphasizing co-creation and mutual benefits that improve the potential for technological and social advancements.

Reflecting on the implications of our findings within the quadruple helix model, it becomes evident that the advanced machine learning techniques utilized in this study enhance retail management and simultaneously foster a collaborative synergistic environment across academia, industry, government, and civil society. Such synergy is essential for driving innovation and socio-economic development, demonstrating the significant benefits of integrating academic research with practical applications across various dimensions of society.

Our collaboration with major retail players in Portugal translates academic research into practical solutions that support government policymaking and provide empirical data on consumer behavior. The societal impact of our study manifests consumer satisfaction and trust, which are crucial for the sustainability of retail businesses and the overall well-being of the community.

The potential for these findings to influence public policy, particularly in consumer protection and business regulation, highlights the importance of ongoing partnerships between academia and governmental bodies. By continuously integrating new insights from sentiment analysis research into policy development, we ensure that regulations evolve in line with technological advancements and market dynamics, ultimately benefiting the broader society.

The future scope of the present research is expansive, aiming to integrate more advanced machine learning architectures like long short-term memory networks and convolutional neural networks. These architectures have demonstrated remarkable efficacy in handling complex textual data and hold the promise of enhancing the predictive accuracy of our models further. Our study contributes to the existing body of knowledge on sentiment analysis within the retail sector, opening avenues for future research in improving model accuracy and adapting pre-trained models to accommodate diverse linguistic and cultural nuances.

**Author Contributions:** Conceptualization, C.C., V.L., A.C.B. and A.F. Data curation, C.A. and C.C. Formal analysis, C.C., V.L., A.C.B. and A.F. Investigation, C.A., C.C. and V.L. Methodology, C.A., C.C., V.L., A.C.B. and A.F. Writing—original draft, C.A., A.C.B. and A.F. Writing—review and editing, V.L. and C.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by FONDECYT, grant number 1200525 (V.L.), from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science, Technology, Knowledge, and Innovation, as well as by Portuguese funds through the CMAT–Research Centre of Mathematics of the University of Minho, within projects UIDB/00013/2020 https://doi.org/10.54499/UIDB/00013/2020 (accessed on 19 May 2024) and UIDP/00013/2020 https://doi.org/10.54499/UIDP/00013/2020 (accessed on 19 May 2024) (C.C.).

**Data Availability Statement:** Data and codes are available from the authors upon request keeping in mind that some specific criteria utilized by the consultancy remain proprietary and confidential.

**Acknowledgments:** The authors would like to thank the editors and reviewers for their constructive comments, which led to improvements in the presentation of the article.

**Conflicts of Interest:** There are no conflicts of interest declared by the authors.

## References

1. Zhang, X.; Tu, Y.; Zhong, K. Are eBay's feedback ratings consistent with the sentiments embedded in textual comments? An empirical study. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 2257–2272. [CrossRef]
2. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*; Pearson: Upper Saddle River, NJ, USA, 2009.
3. Zhang, X.; Guo, F.; Chen, T.; Pan, L.; Beliakov, G.; Wu, J. A brief survey of machine learning and deep learning techniques for e-commerce research. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 2188–2216. [CrossRef]
4. Cavalcante, T.; Ospina, R.; Leiva, V.; Martin-Barreiro, C.; Cabezas, X. Weibull regression and machine learning survival models: Methodology, comparison, and application to biomedical data related to cardiac surgery. *Biology* **2023**, *11*, 1394. [CrossRef] [PubMed]
5. Barros, M.; Galea, M.; Gonzalez, M.; Leiva, V. Influence diagnostics in the tobit censored response model. *Stat. Methods Appl.* **2010**, *19*, 379–397. [CrossRef]
6. Manchini, C.; Ospina, R.; Leiva, V.; Martin-Barreiro, C. A new approach to data differential privacy based on regression models under heteroscedasticity with applications to machine learning repository data. *Inf. Sci.* **2023**, *627*, 280–300. [CrossRef]
7. Sardar, I.; Akbar, M.A.; Leiva, V.; Alsanad, A.; Mishra, P. Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: Methodology, evaluation, and case study in SAARC countries. *Stoch. Environ. Res. Risk Assess.* **2023**, *37*, 345–359. [CrossRef]
8. Wang, Z.; Wang, L.; Ji, Y.; Zuo, L.; Qu, S. A novel data-driven weighted sentiment analysis based on information entropy for perceived satisfaction. *J. Retail. Consum. Serv.* **2022**, *68*, 103038. [CrossRef]
9. Sanchez, L.; Leiva, V.; Galea, M.; Saulo, H. Birnbaum-Saunders quantile regression and its diagnostics with application to economic data. *Appl. Stoch. Model. Bus. Ind.* **2021**, *37*, 53–73.
10. Mowlaei, M.E.; Abadeh, M.S.; Keshavarz, H. Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Syst. Appl.* **2020**, *148*, 113–234. [CrossRef]
11. Kumar, A.; Uma, G. Intelligent sentinet-based lexicon for context-aware sentiment analysis: Optimized neural network for sentiment classification on social media. *J. Supercomput.* **2021**, *77*, 12801–12825. [CrossRef]
12. Ruffer, N.; Knitza, J.; Krusche, M. Covid4Rheum: An analytical twitter study in the time of the COVID-19 pandemic. *Rheumatol. Int.* **2020**, *40*, 2031–2037. [CrossRef] [PubMed]
13. Park, H.W.; Park, S.; Chong, M. Conversations and medical news frames on Twitter: Infodemiological study on COVID-19 in South Korea. *J. Med. Internet Res.* **2020**, *22*, e18897. [CrossRef] [PubMed]
14. Cortis, K.; Davis, R. Over a decade of social opinion mining: A systematic review. Artif. Intell. Rev. **2021**, *54*, 4873–4965. [CrossRef] [PubMed]
15. Arora, A.; Chakraborty, P.; Bhatia, M.; Mittal, P. Role of emotion in excessive use of twitter during COVID-19 imposed lockdown in India. *J. Technol. Behav. Sci.* **2021**, *6*, 370–377. [CrossRef] [PubMed]

16. Ahmad, S.; Asghar, M.Z.; Alotaibi, F.M.; Awan, I. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Hum.-Centric Comput. Inf. Sci.* **2019**, *9*, 27. [CrossRef]
17. Subhashini, L.; Li, Y.; Zhang, J.; Atukorale, A.S.; Wu, Y. Mining and classifying customer reviews: A survey. *Artif. Intell. Rev.* **2021**, *54*, 6343–6389. [CrossRef]
18. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]
19. Stanca, L.; Dabija, D.C.; Câmpian, V. Qualitative analysis of customer behavior in the retail industry during the COVID-19 pandemic: A word-cloud and sentiment analysis approach. *J. Retail. Consum. Serv.* **2023**, *75*, 103543. [CrossRef]
20. Bustos, N.; Tello, M.; Droppelmann, G.; Garcia, N.; Feijoo, F.; Leiva, V. Machine learning techniques as an efficient alternative diagnostic tool for COVID-19 cases. *Signa Vitae* **2022**, *18*, 23–33.
21. Pereira, D.A. A survey of sentiment analysis in the Portuguese language. *Artif. Intell. Rev.* **2021**, *54*, 1087–1115. [CrossRef]
22. Gonçalo Oliveira, H. Automatic generation of creative text in Portuguese: An overview. *Lang. Resour. Eval.* **2024**, *58*, 7–41. [CrossRef] [PubMed]
23. Chaouch, H.; Charfeddine, S.; Aoun, S.B.; Jerbi, H.; Leiva, V. Multiscale monitoring using machine learning methods: New methodology and an industrial application to a photovoltaic system. *Mathematics* **2022**, *10*, 890. [CrossRef]
24. Nor, A.K.M.; Pedapati, S.R.; Muhammad, M.; Leiva, V. Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data. *Mathematics* **2022**, *10*, 554. [CrossRef]
25. Ma, L.; Zhang, Y.; Leiva, V.; Liu, S.; Ma, T. A new clustering algorithm based on a radar scanning strategy with applications to machine learning data. *Expert Syst. Appl.* **2022**, *191*, 116143. [CrossRef]
26. Hassonah, M.A.; Al-Sayyed, R.; Rodan, A.; Al-Zoubi, A.M.; Aljarah, I.; Faris, H. An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter. *Knowl.-Based Syst.* **2020**, *192*, 105–353. [CrossRef]
27. Chang, J.R.; Liang, H.Y.; Chen, L.S.; Chang, C.W. Novel feature selection approaches for improving the performance of sentiment classification. *J. Ambient. Intell. Humaniz. Comput.* **2024**, *in press*. [CrossRef]
28. Taylan, O.; Alkabaa, A.S.; Alqabbaa, H.S.; Pamukcu, E. Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods. *Biology* **2023**, *12*, 1179. [CrossRef] [PubMed]
29. Thet, T.T.; Na, J.C.; Khoo, C.S. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.* **2010**, *36*, 823–848. [CrossRef]
30. Tan, K.L.; Lee, C.P.; Lim, K.M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Appl. Sci.* **2023**, *13*, 4550. [CrossRef]
31. Sharma, P.; Singh, A.K.; Leiva, V.; Cabezas, X.; Martin-Barreiro, C. Modern multivariate statistical methods for evaluating the impact of WhatsApp on academic performance: Methodology and case study in India. *Appl. Sci.* **2022**, *12*, 6141. [CrossRef]
32. Palacios, C.A.; Reyes-Suarez, J.A.; Bearzotti, L.A.; Leiva, V.; Marchant, C. Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy* **2021**, *23*, 485. [CrossRef] [PubMed]
33. Rish, I. An empirical study of the naive Bayes classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 4–6 August 2001; pp. 41–46.
34. Hand, D.J.; Yu, K. Idiot's Bayes: Not so stupid after all? *Int. Stat. Rev.* **2001**, *69*, 385–398.
35. Mitchell, T.M. *Machine Learning*; McGraw Hill: New York, NY, USA, 1997.
36. Lewis, D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 21–23 April 1998; pp. 4–15.
37. Agresti, A. *Analysis of Ordinal Categorical Data*; Wiley: New York, NY, USA, 2010.
38. Hosmer, D.; Lemeshow, S.; Sturdivant, R. *Applied Logistic Regression*; Wiley: New York, NY, USA, 2013.
39. Cox, D.R.; Snell, E.J. *The Analysis of Binary Data*; Chapman and Hall: London, UK, 1989.
40. McFadden, D. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*; Zarembka, P., Ed.; Academic Press: New York, NY, USA, 1974; pp. 105–142.
41. Nagelkerke, N.J.D. A note on a general definition of the coefficient of determination. *Biometrika* **1991**, *78*, 691–692. [CrossRef]
42. Osborne, J. *Best Practices in Logistic Regression*; SAGE Publications: Thousand Oaks, CA, USA, 2017.
43. Pituch, K.; Stevens, J. *Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS*; Taylor and Francis: New York, NY, USA, 2015.
44. Hensher, D.; Stopher, P. *Behavioural Travel Modelling*; Taylor and Francis: New York, NY, USA, 2021.
45. Ashqar, H.; Shaheen, Q.; Ashur, S.; Rakha, H. Impact of risk factors on work zone crashes using logistic models and random forest. In Proceedings of the IEEE Intelligent Transportation Systems Conference, Indianapolis, IN, USA, 19–22 September 2021; pp. 1815–1820.
46. Long, J. *Regression Models for Categorical and Limited Dependent Variables*; SAGE Publications: Thousand Oaks, CA, USA, 1997.
47. McCullagh, P. Regression models for ordinal data. *J. R. Stat. Soc. B* **1980**, *42*, 109–142. [CrossRef]
48. Donaldson, G.W. Ridit scores for analysis and interpretation of ordinal pain data. *Eur. J. Pain* **1998**, *2*, 221–227. [CrossRef] [PubMed]
49. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
50. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
51. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
52. Breiman, L.; Cutler, R.A. Random forests machine learning. *J. Clin. Microbiol.* **2001**, *2*, 199–228.

53. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13–17 August 2016; pp. 785–794.
54. Akritidis, L.; Bozanis, P. How dimensionality reduction affects sentiment analysis NLP tasks: An experimental study. In Proceedings of the 18th International Conference on Artificial Intelligence Applications and Innovations, Crete, Greece, 17–20 June 2022; pp. 301–312.
55. Akritidis, L.; Bozanis, P. Low dimensional text representations for sentiment analysis NLP tasks. *SN Comput. Sci.* **2023**, *4*, 474. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.