Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 5** : 1–12.
ISSN-0037-8712
doi:10.5269/bspm.80619

Classification of Unstructured Text for RPA: A Comparative Machine Learning Analysis

Zeynep Orpek and Adem Orpek and Onder Sahinaslan

ABSTRACT: Effective human-robot collaboration in digital work environments depends on automated systems correctly interpreting human-generated free-text. Unstructured text types such as bug reports, customer requests, call center records, emails, and free comment fields cannot be processed directly by robotic process automation (RPA) due to the contextual expressions, typos, and stylistic inconsistencies they contain. RPA processes operate rule based. Making sense of this unstructured data, which does not conform to the rules, will strengthen human-robot interaction. In this study, it is proposed to use machine learning methods to provide automatic classification of unstructured texts. In this study, five basic classification algorithms (Logistic Regression, Naive Bayes, Support Vector Machines, Decision Trees, and Random Forest) frequently used in the literature were compared in terms of their features and computational methods. The findings reveal that classifying unstructured text has the potential to significantly increase the accuracy of RPA-based workflows. In this way, it is anticipated that improvements can be made by shortening process times through the interpretation of texts in RPA processes. As a result, it is envisaged that robot-human interaction can become more efficient by making sense of the unstructured data in RPA processes.

Keywords: Robotic Process Automation (RPA), unstructured data, text classification, machine learning.

Contents

1. Introduction

RPA is defined as a business process automation approach that automates repetitive and rule-based tasks that employees previously performed manually through software robots [1]. In order for RPA systems to operate reliably, the workflows they execute must be based on clearly defined, consistent, and structured data inputs by users. However, the data types encountered in real-world business environments often fail to meet these requirements. Text-based data, especially those such as bug reports, customer requests, call center logs, emails, online forms, or user-written free-form comments, is inherently unstructured. Unstructured data includes data types that do not have a specific data model or schema, are not formally regular, and are not directly suitable to analysis [2]. This data heavily contains context-dependent statements, typographical errors, missing or redundant information, and inconsistencies. This makes it difficult for robots to accurately determine which workflow to initiate next and increases the probability of errors in the process.

RPA works on the principle of imitating the steps a human performs on a computer by software robots. The robot reads the screen, repeats click, enters data into forms, switches to a different workflow when a certain condition is met, and performs all of these operations without error within the framework of defined rules. Just as a human opens an email, reads its content, and directs it to the relevant department, the robot similarly receives data. However, the robot cannot interpret the text itself. In other words, the robot knows what to do, but it cannot understand the text it receives. RPA approaches only work effectively on structured data such as tables, forms, and databases [3]. Therefore, for robots

2020 *Mathematics Subject Classification:* 68T07, 68T05, 68T45.
Submitted December 05, 2025. Published February 14, 2026

to make correct decisions, the input must be defined structurally. Machine learning is a technology that mimics human intelligence using deep learning algorithms [4]. At this point, machine learning models can come into play and classify the free-text that the robot cannot understand and enabling the robot to understand unstructured texts.

In this context, the classification of unstructured texts has become not only a supporting component for the success of RPA systems but also a critical requirement for the reliable execution of human-robot collaboration. In this study, five basic machine learning algorithms widely used in the literature were selected, and a comparative evaluation was carried out by examining the mathematical foundations and computational processes of these models.

2. Literature Review

This study examines the definition of unstructured data and the application areas of RPA. Furthermore, existing research on five widely used classification models: Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Decision Trees (DT), and Naive Bayes (NB) is comprehensively reviewed.

Dolgun et al. (2009) classify data sources into two categories: structured and unstructured data. Structured data refers to data types that are organized according to a specific schema and can be directly processed using query languages like SQL in relational databases. These data allow the direct application of classical statistical analysis methods and data mining algorithms thanks to their predefined properties such as column, row, and data type. In their study, they indicate that approximately 80% of the data generated and stored in organizations is unstructured. Therefore, they emphasize that an additional transformation step is required for this data type to be usable in analytical processes [2]. As stated in this study, incorporating unstructured data into analytical models will only be possible by first converting the content into structural form through methods such as text mining and web mining.

Eberendu (2016) highlights the increasing amount of unstructured data, emphasizing that this data type has become critical information for corporate decision-making processes. The study states that unstructured data generated from various sources, such as emails and multimedia content, cannot be processed with traditional models, complicating analysis and classification processes. It is noted that 80–85% of existing data exists in unstructured formats, yet this data cannot be directly transferred to operational decisions without being interpreted. This study shows that text-heavy data in particular plays a critical role in generating corporate insights and that such content cannot be used effectively in business processes without being converted into structured information. The need to classify unstructured text data generated throughout the process in organizations and transform it into usable structured decision inputs for RPA is identified as a fundamental problem [5].

Atalay and Çelik (2017) examine the concept of big data in terms of volume, velocity, and diversity, and discuss the decisive role of artificial intelligence and machine learning techniques in extracting meaningful and high-value-added information from this data. The study provides examples of how methods such as classification, artificial neural networks, text and web mining, and sentiment analysis are applied to large, often unstructured, data sources. It emphasizes the inadequacy of traditional methods in the face of this volume and diversity of data. They state that data mining and machine learning-based approaches have become essential, particularly in the analysis of text-based and dynamic data sources and present a broad framework for the relationship between big data analytics and artificial intelligence techniques [6].

Dalsaniya (2022), focusing on the concept of cognitive robotic process automation, emphasizes that traditional rule-based RPA approaches are only effective on structured data such as tables, forms, and databases, whereas unstructured sources such as email, scanned documents, and free-text, which constitute the majority of corporate data, cannot be adequately processed within this framework. The study emphasizes that by adding NLP, OCR, and machine learning components to the RPA architecture, cognitive RPA systems can automate text and document-based processes in sectors such as healthcare, finance, law, and manufacturing. Case studies demonstrate that this results in significant improvements in processing speed, accuracy, cost savings, and employee productivity [3].

Human–robot interaction (HRI) is positioned as a fundamental research area that addresses the cognitive and communicative processes necessary for robotic systems to operate safely and understandably in human-centered environments. Murphy et al. (2010) emphasize that HRI is not limited to physical

interaction. They state that it encompasses multidimensional elements such as shared perception, joint attention, communication modes, decision-making mechanisms, and task coordination. This multi-layered structure makes it necessary for robots to be able to correctly interpret human behavior, intentions, and textual or verbal commands. This framework of HRI in the literature emphasizes that robots should move from being merely mechanical systems based on certain rules to cognitive actors that can make sense of human-generated data [7].

Madakam et al. (2019) study positions RPA as one of the core components of the digital workforce and discuss in a broad framework how this technology transforms organizational functioning. They state that RPA not only speeds up individual tasks but also reduces costs for businesses and creates 24/7 processes thanks to the "digital workforce" it provides. Therefore, they state that RPA has become a competitive advantage in many sectors, such as banking, insurance, retail, and energy. The study examines RPA as a platform that can integrate with technologies such as artificial intelligence, machine learning, deep learning, data analytics, virtual reality, and blockchain [1].

Yarlagadda (2018) examines the transformation of RPA and AI-based automation in business processes, emphasizing the high speed and accuracy with which repetitive tasks are performed through software robots, particularly in the finance and service sectors. The study examines RPA's potential to reduce costs, increase operational efficiency, and minimize human error. It is noted that this type of automation is largely limited to structured data and rule-based processes. According to Yarlagadda, while RPA excels at operations based on explicit rules, such as data entry, system integration, and scheduled transactions, it cannot directly process unstructured process data, such as textual user requests, email content, and customer feedback [8]. These findings reinforce the motivation for this study, confirming the view that current RPA solutions cannot make sense of unstructured data and, therefore, the need for hybrid approaches supported by more advanced machine learning methods.

Göker and Tekedere (2017) conducted a study that automatically evaluated opinions shared online regarding the Fatih Project in the field of education using text mining methods. In this study, software was developed to analyze unstructured text obtained from forums and similar sources. This software transformed the texts into a structured dataset by processing them through the steps of transformation, stop word removal, and stemming. The performances of different machine learning algorithms were then compared on 444 documents containing positive/negative opinions, and the Sequential Minimal Optimization (SMO) algorithm reported the highest accuracy rate of 88.73% [9]. In this respect, this study parallels the approach used to classify free-text fields in RPA processes, as it converts Turkish unstructured texts into numerical features using a standard pre-processing pipeline and demonstrates that these features can be successfully classified with supervised learning algorithms.

Başkaya and Aydın (2017) used text mining techniques to automatically classify Turkish news texts in the categories of economy, politics, sports, and health. Data cleaning, stop word removal, and stemming using Zemberek were performed on 80 news texts. Root word and bigram representations were used for feature extraction. NB, SVM, J48, and FR models were compared on feature vectors generated using TF and TF-IDF weighting methods. The highest accuracy was achieved with the combination of root word + TF weighting + RF [10]. The study demonstrates that Turkish text classification can achieve high success with appropriate feature representation and feature selection.

A systematic literature review by Baviskar et al. (2023) comprehensively evaluates current methods for automatically processing unstructured documents using artificial intelligence techniques. The study compares the performance of classical machine learning and deep learning models in text-based organizational data classification, information extraction, and document understanding. The integration of these methods with process automation and decision support systems is discussed. It emphasizes that the process of transforming unstructured data into structured decisions is critical for organizational automation [11].

Dreiseitl and Ohno-Machado (2002) compare LR and artificial neural networks in the context of medical data classification problems, discussing their similarities and differences in terms of their statistical pattern recognition roots, model-building steps, and evaluation criteria. The study examines LR and artificial neural network models alongside other classification algorithms such as k-nearest neighbor, DT, and SVM. Emphasis is placed on preventing overfitting, parametric/semi-parametric architecture, variable selection, cross-validation, and bootstrap. A review of 72 medical studies demonstrates that, in

most cases, there is no clear superiority between the different models, and that model quality is strongly dependent on the dataset used, the selected parameters, and the accurately reported evaluation criteria [12].

A study by LaValley (2008) comprehensively summarizes how logistic regression, a widely used method for analyzing two-class outcomes, is applied in research. The article explains why this method is frequently preferred, particularly in medical research, in which situations it yields more accurate results, and how the model is evaluated. The author emphasizes that logistic regression holds a significant place in scientific research due to its ability to analyze multiple variables simultaneously, its ability to understand relationships between samples with different characteristics, and its reliability as a method for classification studies [13]. In these respects, the study clearly explains the application logic of basic machine learning approaches used in classification problems.

Venigandla (2022) argues that RPA provides high-quality data to AI models by automating repetitive data collection and pre-processing tasks and demonstrates that AI/ML models support complex decision-making processes by extracting meaningful patterns from this data. The study focuses specifically on standardizing unstructured medical data using RPA and improving diagnostic accuracy by classifying it with ML algorithms. While the industry is diverse, the study provides an important conceptual foundation for this paper regarding the automatic processing, classification, and transfer of unstructured text data to RPA as decision input [14].

Amarappa and Sathyanarayana (2013) describe SVM as a powerful machine learning method developed for binary classification problems, capable of modeling both linear and nonlinear distinctions using kernel functions. The core idea of SVM is to separate training examples with a hyperplane that best separates classes and keeps the margin (inter-class gap) as wide as possible. Using only the examples closest to the boundary, namely support vectors, when determining the decision surface allows the method to demonstrate effective and generalizable classification performance even in high-dimensional feature spaces [15].

Tang et al. (2009) investigated different rebalancing strategies to improve the performance of SVM on highly imbalanced datasets. They compared cost-sensitive learning, minority class oversampling, majority class under-sampling, and the granular SVM-based GSVM-RU approach. They evaluated the results using metrics such as G-mean, AUC-ROC, F-measure, and AUC-PR [16]. The results demonstrate the importance of sampling and cost-sensitive models for accurate prediction of rare classes in imbalanced data structures.

Despite relying on the assumption that features are conditionally independent of class, the NB classifier performs better than expected in many practical problems. Rish (2001) systematically examined the data characteristics that affect NB's performance using Monte Carlo simulations, demonstrating that classification error decreases, particularly in low-entropy distributions and in extreme cases where features are completely independent or functionally interconnected. The study also emphasizes that model error is better explained not only by the strength of feature dependencies but also by the amount of information lost about the class due to the independence assumption [17].

Although the NB classifier is widely used in text classification problems due to its simplicity and computational efficiency, it can cause performance degradation, especially in multi-class and imbalanced datasets, due to structural constraints such as the independence assumption and zero frequency. The Naive Bayes Enrichment Method (NBEM), proposed by Peretz et al. (2024), offers an ensemble approach that combines multiple NB variants with weights suitable for different distributions. This method provides significant improvements in recall and F1 scores. The study demonstrates that feature weighting and sub-setting strategies significantly improve NB performance, especially in heterogeneous data types [18]. These findings suggest that NB-based approaches, when appropriately developed, can be an effective alternative for classifying unstructured, complex data types.

Desai et al. (2021) developed an artificial intelligence-supported invoice processing system to facilitate the processing of unstructured documents in RPA processes. The study used the Intelligent Document Processing (IDP) approach to extract text from PDF and scanned documents, perform field classification, and perform data validation. This study demonstrates that NB-based text classification models, when integrated with RPA, automatically convert unstructured data into a structured form, resulting in significant efficiency gains in workflows [19]. While the application area is different, this study is instructive

for this research by demonstrating that categorizing unstructured process data with machine learning models strengthens decision support mechanisms in RPA scenarios.

DT is a machine learning algorithm used in both classification and regression problems within the context of supervised learning, hierarchically partitioning data according to decision rules. Starting from the root node, the model performs sequential partitions aimed at dividing the dataset into the most homogeneous subgroups. Each internal node represents a decision rule for a feature, while leaf nodes represent the predicted class or value. Bansal et al. (2022) state that decision trees are particularly advantageous in terms of interpretability, do not require additional assumptions about the data type, and can work flexibly with both numerical and categorical variables. However, significant limitations include being prone to overfitting in irregular or multivariate datasets and the potential for information loss due to processing continuous variables by dividing them into classes. The model's performance is shaped by feature selection criteria such as information gain and the Gini index, and over-branching problems can be significantly mitigated when appropriate pruning techniques are applied. With these aspects, DT is a frequently preferred method in applications where nonlinear patterns need to be explained and decision logic needs to be clearly monitored [20].

RF is an ensemble classifier proposed by Breiman (2001) that performs predictions by combining multiple decision trees. Each tree is trained on a random subset of training samples drawn from the bootstrap method, and a split is performed using a limited number of randomly selected features at each node. This structure reduces model variance compared to a single deep decision tree, significantly reducing the risk of overfitting [21].

Belgiu and Drăguț (2016) emphasize that RF yields successful classification results, particularly on high-dimensional and multi-collinear data, is relatively fast in computation, and is generally insensitive to overfitting. Their study also notes that RF's inherently generated out-of-bag error estimate and variable importance allow for both reliable assessment of model accuracy and the selection of the most discriminative features for classification [22]. These features make RF a powerful and practical machine learning approach.

The reviewed literature demonstrates that unstructured data plays a critical role in enterprise knowledge generation and process automation. However, it reveals that this data often cannot be directly processed by traditional RPA solutions. Studies demonstrate that unstructured text can be converted into structured decisions using text mining and machine learning techniques, and that this transformation provides significant performance gains, particularly in decision support systems and cognitive RPA applications. Significant gaps exist in the literature regarding the systematic classification of process data generated within the context of RPA technology, the comparison of different machine learning algorithms on the same problem, and the holistic consideration of the mathematical foundations of these models. This study aims to comparatively evaluate the features and classification performance of LR, NB, SVM, DT, and RF algorithms.

3. Methodology

In this study, five of the most used algorithms in the literature were chosen for the classification of unstructured text: LR, NB, SVM, DT, and RF. These algorithms are widely used in a wide variety of fields, including medicine, finance, marketing, fraud detection, spam filtering, sentiment analysis, and bug classification, and are particularly effective for problems with big data and sparse feature spaces, such as text data. In the context of this study, these models were used to assign free-text data to appropriate categories to enhance human-robot collaboration. Thus, each classification decision becomes a decision input that determines which workflow will be triggered by the robotic processes running in the background.

Logistic Regression

LR is a basic statistical model used in binary classification problems. Estimates the probability that a categorical dependent variable belongs to a particular class based on one or more independent variables. Unlike linear regression, which produces continuous outputs, this method is particularly suitable for modeling binary outcomes such as yes/no or true/false. The logistic (sigmoid) function underlying the model enables the classification process by transforming any real-valued input to produce a probability

value between 0 and 1. As outlined in Özkul's (2024) study, the mathematical basis of logistic regression is expressed as follows, using the logit transformation of probability [23].

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{3.1}$$

In this equation, p represents the probability of the event occurring, while $\beta_0, \beta_1, \ldots, \beta_k$ represents the model coefficients. Similarly, $X_1, X_2, \ldots, X_k$ represent independent variables. LR uses the maximum likelihood method to obtain the parameters that provide the highest fit to the observed data and, in this respect, offers a powerful estimation tool for binary categorization problems.

LR has a wide range of applications and is widely used in decision support processes across various disciplines. In medicine, it produces effective results in binary decision problems such as disease presence, treatment success, or identifying at-risk patient groups. In the financial sector, it is frequently used for classification-based analyses such as credit risk scoring, customer return (churn) prediction, and fraud detection. LR's ability to generate probabilistic outputs has the potential to provide advantages in tasks such as automatically assigning free-text or unstructured input to specific classes and process routing in RPA processes. Thanks to the model's interpretable structure, automated decision rules based on specific constraints can be easily created in RPA scenarios requiring trigger conditions. This optimizes processes in terms of both speed and accuracy. In this respect, LR stands out as a reliable and effective method for RPA-based text classification and workflow automation.

Naive Bayes

NB classifier is an algorithm in the supervised learning class that is widely used in fields such as data mining, machine learning, and sentiment analysis [24]. They perform particularly well in tasks requiring rapid parsing of high-dimensional text data, such as email filtering, spam detection, sentiment analysis, news classification, and document categorization. They are also frequently used in applications such as risk prediction, anomaly detection, and early warning systems in medicine, finance, and security.

In the NB classifier, after estimating the model parameters from the training documents, the probability that a new test document d_i belongs to each class c_j is calculated using Bayes' rule. McCallum and Nigam (1998) formulate this classification step as follows [25].

$$P(c_j \mid d_i; \theta) = \frac{P(c_j \mid \theta)\, P(d_i \mid c_j; \theta_j)}{P(d_i \mid \theta)} \tag{3.2}$$

In this equation $P(c_j \mid d_i; \theta)$, represents the probability that the test document belongs to class c_j (posterior). $P(c_j \mid \theta)$. represents the prior probability of the class obtained from the training data (prior). $P(d_i \mid c_j; \theta_j)$, represents the likelihood of the document appearing in the relevant class (probability). $P(d_i \mid \theta)$ represents the probability of observing the document across all classes (evidence). Classification decisions are made by calculating this posterior probability for each class and selecting the class with the highest value.

In the RPA context, NB's key advantage is its ability to quickly classify unstructured text input with low computational cost. This allows for the reliable automation of text-based decision steps in RPA workflows, such as categorizing customer requests, prioritizing bug logs, routing support requests, and distinguishing exception scenarios. The model's simple yet effective structure can facilitate real-time, scalable operation of RPA robots, particularly in processes with large text volumes, and increase process efficiency.

Support Vector Machine

SVM is widely used in many fields such as text classification, sentiment analysis, document categorization, bio-informatics, image recognition, and financial risk analysis, thanks to its structure that can create strong separation boundaries in high-dimensional data spaces. SVM is particularly successful in datasets with complex patterns due to its ability to produce effective decision boundaries in nonlinear data via kernel functions.

In SVM, the goal is to obtain an optimal decision boundary that minimizes classification errors while maximizing the separation margin between classes. Jakkula (2006) defines the basic optimization problem of soft–margin SVM as follows [26]:

$$\begin{aligned} \min_{w,\,b,\,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0 \end{aligned} \tag{3.3}$$

In this formula, w represents the weight vector that determines the direction of the decision boundary. b represents the bias term that determines the location of the hyperplane. x_i represents each training example, and $y_i \in \{-1, +1\}$ represents the class labels of these examples. ξ_i are slack variables and represent the magnitude of the margin violation or error amount for each example. The parameter C balances the width of the margin and the penalty given to classification errors. Large values of C penalize errors more, while small values allow for a wider margin. ℓ represents the total number of training examples. This structure forms the basis of SVM's flexible and robust classification capacity, which both controls errors and provides a maximum margin.

In the context of RPA, SVM can achieve high accuracy in tasks such as classifying free-text fields, categorizing customer messages, estimating the criticality of bug records, and automatically parsing process exceptions. Its margin-maximizing structure helps RPA robots make more accurate decisions by highlighting subtle distinctions between classes in text. In this respect, SVM can offer a reliable and effective method for accuracy-focused classification models in RPA workflows.

Decision Tree

DT is used in a wide range of applications, including diagnostic support systems in healthcare, credit evaluation and fraud detection in finance, customer segmentation and targeting in marketing, success prediction in education, and quality control in industry. The model's rule-based structure offers a significant advantage, particularly in decision-making processes requiring explainability.

DT are intuitive and interpretable machine learning methods widely used in classification and prediction problems, representing relationships between variables through a branched structure. Song and Lu (2015) define DTs as a hierarchical structure consisting of a root node, internal nodes, and leaf nodes. They emphasize that this method is particularly effective for large, complex, or non-parametric datasets. The DT model creates a chain of decision rules by repeatedly splitting inputs to maximize homogeneity of the target variable (e.g., class label). Each split is determined by the variable that provides the highest information gain or purity increase; thus, the model generates increasingly purer subgroups from the root node to the leaves [27]. As a result, DT functions both as a model that performs classification and as a visualization tool that describes the data structure.

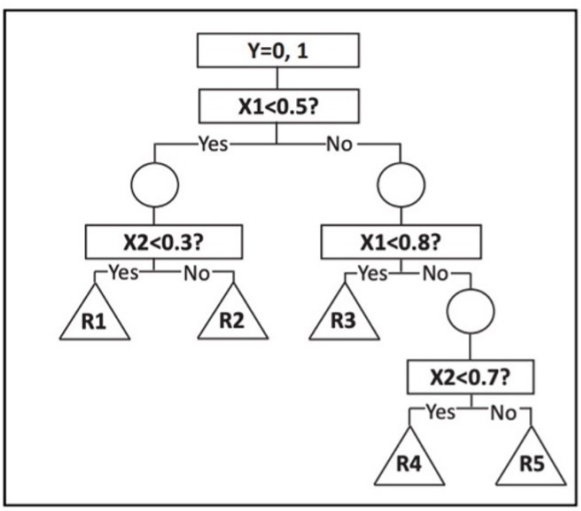

Figure 1: An Example DT Classification Model [27].

Figure 1, presented in Song and Lu's (2015) study, shows an example DT constructed using a binary

target variable (Y = 0 or 1) and two continuous variables (X1 and X2). In this example model, the root node represents the variable Y, and the initial split is made based on a threshold of 0.5 for X1. Following this split, the tree is split into two branches, and each branch is further split based on different thresholds for variables X1 and X2. Each path ends with a leaf node (R1–R5), and these leaves represent the final classification results, representing the homogeneous class distribution of observations that meet the given conditions. Each path in the figure also corresponds to decision rules that can be interpreted as "if–then"; for example, "If X1 ¡ 0.5 and X2 ¡ 0.3, class is assigned to R1." This structure clearly demonstrates how the DT branches, which variables are important in which order, and which values are used to perform the classification. This figure and explanation contribute to the methodology section by showing the basic working principles of DT, division logic, variable selection processes, and how the results are interpreted.

Within the RPA context, DT can be used for tasks such as classifying customer requests, identifying process exceptions, and defining automated routing steps in workflows. The tree structure's decision rules can produce clear and enforceable rules that RPA robots can use directly. This can achieve both interpretability and operational efficiency in the automation of text-based processes.

Random Forest

RF is an ensemble learning method based on the creation of a large number of decision trees during the training phase and the aggregation of their outputs. In classification problems, the final decision is determined by the majority vote (mode) of individual trees. In regression problems, it is obtained by averaging the tree predictions. This approach significantly reduces the tendency of a single decision tree to overfit the training data because the model utilizes the diversity of numerous trees grown on different training subsets created using the bootstrap method [21]. Furthermore, one of the key outputs of RF models is feature importance scores. These scores contribute to the understanding of the deterministic features of the model by indicating which variables are used more frequently and effectively in the decision structures of the trees in the forest [28].

Within the context of RPA, RF can offer significant advantages, particularly in tasks requiring high accuracy, such as classifying free-text fields, parsing process exceptions, and automatically categorizing customer requests. The unanimous decision-making of multiple trees can provide RPA robots with more reliable output. This can reduce the risk of bug in text-based automation processes and increase operational efficiency.

Comparison of Classification Algorithms

This section examines the conceptual structures, application areas, and methodological features of the main classification algorithms used in this study within a comparative framework. Different classification algorithms offer varying advantages and limitations depending on data type, sample size, feature size, and problem structure. Therefore, evaluating the theoretical foundations and performance dynamics of these methods during the model selection process is critical for selecting the right algorithm for classification problems.

In this context, the five core machine learning algorithms used in this study (LR, NB, SVM, DT, and RF) were evaluated based on Osisanwo et al.'s (2017) comparative study of supervised learning algorithms. This study systematically compares different algorithms on both small and large datasets based on metrics such as accuracy, kappa statistics, MAE (Mean Absolute Error), accuracy rate, misclassification rate, and model building time. Table 1 summarizes the main features of the algorithms, their advantages and disadvantages, and the experimental findings of related work [29].

This comparison allows us to evaluate the behavior of the algorithms in different data scenarios, both theoretically and experimentally. According to the findings of the study, SVM stands out as the method that exhibits the highest accuracy on both large and small datasets. NB and RF demonstrate competitive performance, especially on larger datasets. DT and LR achieve more limited accuracy on comprehensive datasets. These results suggest that the selection of a classification algorithm should be based on the characteristics, distribution, size, and problem context of the dataset rather than a single criterion.

4. Findings and Discussion

The classification algorithms discussed in this study were comparatively evaluated for unstructured text classification problems in the RPA context, based on both experimental results reported in the literature and their mathematical properties. The results of the study indicate that methods capable of

Table 1: Comparison of Used Classification Algorithms

Algorithm	**Key Features**	**Advantages**	**Disadvantages**	**Experimental Findings of the Relevant Study**
LR	The linear decision boundary transforms the weighted sum of the inputs into a probability with a sigmoid.	Interpretable. Fast and cost-effective. Effective for large text sizes.	Limited success with nonlinear data structures.	In terms of accuracy, SVM lags behind NB and RF.
NB	Probabilistic classifier that assumes features are independent of each other.	Very fast. High accuracy on small data sets. Low memory usage.	The assumption of independence is not always realistic.	Ranked 2nd in the large dataset (76.3%). Ranked 3rd in the small dataset.
SVM	Classification based on maximum margin; nonlinear separations with kernel functions if necessary.	Highest accuracy. Robust on high-dimensional data. High generalization capability.	Sensitive to parameter adjustments. Training time may increase with large datasets.	Highest accuracy on both large and small datasets (77.34% and 72.92%).
DT	A tree structure that classifies data by dividing it into branches based on features.	Interpretable. Fast training.	Prone to overlearning. Unstable on data changes.	Low accuracy on both datasets. SVM lags behind NB and RF.
RF	Combination of bootstrap and multiple decision trees with voting.	Resistant to overfitting. High accuracy. Can measure variable importance.	The training time is longer than that of a single tree.	Ranked 3rd in the large dataset (74.7%). Ranked 2nd in the small dataset (71.88%).

modeling nonlinear decision boundaries and with high generalization capacity (such as SVM and RF) stand out, especially in text data with high dimensions and sparse feature spaces. Osisanwo et al. (2017) reported that SVM generally achieved the highest accuracy values on different datasets, while RF and NB provided competitive and stable performance [29]. This result shows that SVM and RF-based approaches are strong candidates for text classification models to be used in RPA processes to support human-robot collaboration.

Despite its simple assumptions, the NB classifier can deliver remarkably successful results in text classification problems. The literature reviewed in this study demonstrates that this model can be used as a powerful baseline method, particularly on small and medium-sized datasets, thanks to its low computational cost and ease of implementation. However, performance can be limited in scenarios where the assumption of independence between features is unrealistic or where inter-class imbalances are significant, and therefore, it may be necessary to augment NB-based models with sampling, weighting, or ensemble structures. In the context of RPA, this suggests that NB can be used as a fast and explainable initial filtering model for high-dimensional process data, while more critical decisions can be delegated to more complex models such as SVM or RF.

While LR and DT don't always deliver the best performance in terms of accuracy, their interpretability and transparency of decision-making mechanisms allow them to play an important complementary role

in RPA scenarios. LR offers advantages in policy definition and threshold setting (e.g., "trigger robot process if probability is above x%) by directly generating class probabilities. DT increases the confidence of process owners by generating if-then rules that human experts can easily read and verify. However, results reported in the literature indicate that these models tend to achieve lower accuracy rates compared to SVM and RF when used alone, especially for complex, nonlinear, and noisy data structures. Therefore, these algorithms are more suitable in most cases, either as components of more advanced ensemble methods or as decision support/interpretation layers.

RF is generally recognized in the literature for its high accuracy, relative resistance to overfitting, and ability to measure variable importance. Especially in text mining applications, when working with high-dimensional feature representations like TF-IDF, RF's ability to both maintain model performance and highlight important features offers a critical advantage in understanding which words or phrases trigger specific workflows in RPA processes. In this respect, RF can be considered not only a classification tool but also an insight-generating component for process analysis and improvement. However, the large number of trees in the model necessitates careful management of computational costs in real-time or resource-constrained environments. The resulting holistic assessment demonstrates that text classification architectures for RPA processes should be based on layered or hybrid structures that leverage the strengths of different algorithms, rather than relying on a single "best" model. For example, NB can provide rapid initial classification, and ambiguous or critical examples can be directed to more powerful models such as SVM or RF. LR can determine probability-based thresholds. The rule and variable importance produced by DT and RF can both provide explainability to process owners and provide input to policy designs that improve human-robot collaboration. Such an approach provides a powerful framework for systematically classifying unstructured texts and integrating them into RPA workflows, enabling robots to make more reliable text-based decisions.

In conclusion, when the literature findings and the mathematical properties of the algorithms are evaluated together, it is seen that in RPA scenarios that aim to enhance human-robot collaboration through the classification of unstructured process data, SVM and RF play complementary roles in terms of their high accuracy and generalization capacity, NB in terms of its speed and scalability advantages, and LR and DT in terms of their interpretability and ease of policy definition. By revealing these multidimensional properties of different classification algorithms, this study provides a guiding discussion on model selection, architectural design, and the performance/interpretability tradeoff for future applied RPA projects.

5. Conclusions and Future Study

Conclusion

In this study, a theoretical framework for the classification of unstructured texts commonly encountered in RPA processes is presented. Five basic machine learning algorithms (LR, NB, SVM, DT, and RF) frequently used in the literature were comparatively examined in terms of their mathematical foundations, methodological features, and reported performance results. The findings demonstrate that the traditional rule-based architecture of RPA cannot handle free-text process data such as bug reports, customer requests, and email content on its own. Therefore, classification models that can transform text into structured decisions play a critical role in ensuring the continuity and reliability of human-robot collaboration. The comparative evaluation revealed that SVM and RF algorithms stand out in terms of accuracy and generalization, particularly for text data with high-dimensional and sparse feature spaces. NB can be used as a powerful baseline model thanks to its speed and computational efficiency, while LR and DT serve as complementary models in terms of interpretability and ease of policy definition. These results indicate that layered or hybrid architectures that leverage the strengths of different models are more suitable for RPA scenarios than a single best algorithm. For example, NB can perform rapid pre-classification and direct uncertain examples to more complex models such as SVM or RF. LR can define threshold-based decision rules and provide explainability of the rules generated by DT/RF to process owners. These strategies can be considered in this context. In general, systematic classification of unstructured texts with machine learning methods enables RPA systems to undertake not only tasks based on rule-based and structured inputs but also decision processes that require content and context. In this respect, the study provides a guiding theoretical foundation for incorporating free-text-based process

data into automation through hybrid architectures enriched with classification algorithms, both in the RPA and human-robot collaboration literatures.

Future Work and Recommendations

Since this study provides a theoretical framework, it is important for future research to evaluate the practical performance of classification models by applying them to real institutional data sets. Using high-dimensional and diverse unstructured text sources, such as bug reports, customer requests, and email content, will provide a clearer picture of the models' behavior in real-world workflows. Additionally, in text-based processes where class imbalance is common, testing sampling methods or cost-sensitive learning strategies can contribute to improving classification accuracy. In future studies, comparing algorithms on text data produced in different sectors will provide a broader evaluation opportunity in terms of model selection. In addition, practical investigation of the integration of classification models into RPA workflows stands out as an important research area that will increase the decision-making capacity of robots, strengthen human-robot collaboration, and improve the quality of automation.

Acknowledgements

The authors acknowledge that some of the findings presented in this paper were previously shared at the 9th International Conference on Mathematical Sciences (ICMS 2025) held at Maltepe University, Istanbul, Türkiye.

References

[1] S. Madakam, R. M. Holmukhe, and D. K. Jaiswal., *The Future Digital Work Force: Robotic Process Automation (RPA). JISTEM - Journal of Information Systems and Technology Management*, vol. 16, e201916001, (2019). doi: 10.4301/S1807-1775201916001

[2] M. Ö. Dolgun, T. G. Özdemir, and D. Oğuz., *Analysis of Unstructured Data in Data Mining: Text and Web Mining. Journal of Statisticians: Statistics and Actuarial Sciences*, vol. 2, no. 2, pp. 48–58, (2009).

[3] N. A. Dalsaniya, *Cognitive Robotic Process Automation (RPA) for Processing Unstructured Data. International Journal of Science and Research Archive*, vol. 7, no. 2, pp. 639–643, (2022).

[4] E. Şahinaslan, M. Günerkan, and Ö. Şahinaslan., *An Alternative Solution Method for the Use of Categorical Data Encoding Technique in Machine Learning. Journal of Intelligent Systems: Theory and Applications*, vol. 6, no. 1, pp. 1–11, 2023, doi:10.38016/jista.1140499.

[5] A. C. Eberendu, *Unstructured Data: An overview of the data of Big Data. International Journal of Computer Trends and Technology*, vol. 38, no. 1, pp. 46–50, (2016).

[6] M. Atalay and E. Çelik, *Artificial intelligence and machine learning applications in big data analysis. Mehmet Akif Ersoy University Journal of Social Sciences Institute*, vol. 9, no. 22, pp. 155–172, (2017). doi: 10.20875/makusobed.309727

[7] R. R. Murphy, T. Nomura, A. Billard, and J. L. Burke., *Human–robot interaction. IEEE Robotics & Automation Magazine*, vol. 17, no. 2, pp. 85–89, (2010).

[8] R. T. Yarlagadda., *The RPA and AI Automation. International Journal of Creative Research Thoughts (IJCRT)*, ISSN 2320-2882, (2018).

[9] H. Göker and H. Tekedere., *Automatic Evaluation of Opinions on the FATİH Project Using Text Mining Methods. Journal of Information Technologies*, vol. 10, no. 3, pp. 291–299, (2017). doi:10.17671/gazibtd.331041

[10] F. Başkaya and İ. Aydın., *Classification of News Texts Using Different Text Mining Methods.* In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–5, IEEE, (2017).

[11] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha., *Efficient Automated Processing of Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions. IEEE Access*, vol. 9, pp. 72894–72936, (2021).

[12] S. Dreiseitl and L. Ohno-Machado., *Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. Journal of Biomedical Informatics*, vol. 35, no. 5–6, pp. 352–359, (2002).

[13] M. P. LaValley., *Logistic Regression. Circulation*, vol. 117, no. 18, pp. 2395–2399, (2008).

[14] K. Venigandla., *Integrating RPA with AI and ML for Enhanced Diagnostic Accuracy in Healthcare. Power System Technology*, vol. 46, no. 4, pp. 33–42, (2022).

[15] S. Amarappa and S. V. Sathyanarayana., *Data Classification Using Support Vector Machine (SVM): A Simplified Approach. International Journal of Electronics and Computer Science Engineering*, vol. 3, pp. 435–445, (2014).

[16] Y. Tang, Y. Q. Zhang, N. V. Chawla, and S. Krasser., *SVMs Modeling for Highly Imbalanced Classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, (2008).

[17] I. Rish., *An Empirical Study of the Naive Bayes Classifier.* In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, no. 22, pp. 41–46, (2001).

[18] O. Peretz, M. Koren, and O. Koren., *Naive Bayes Classifier – An Ensemble Procedure for Recall and Precision Enrichment. Engineering Applications of Artificial Intelligence*, vol. 136, 108972, (2024).

[19] D. Desai, A. Jain, D. Naik, N. Panchal, and D. Sawant., *Invoice Processing Using RPA & AI.* In *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*, (May 2021).

[20] M. Bansal, A. Goyal, and A. Choudhary., *A Comparative Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short-Term Memory Algorithms in Machine Learning. Decision Analytics Journal*, vol. 3, p. 100071, (2022).

[21] L. Breiman., *Random Forests. Machine Learning*, vol. 45, no. 1, pp. 5–32, (2001). doi:10.1023/A:1010933404324

[22] M. Belgiu and L. Drăguţ., *Random Forest in Remote Sensing: A Review of Applications and Future Directions. ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, (2016).

[23] E. Özkul., *Regression and Classification Algorithms in the Analysis of Educational Data: Modeling Student Performance.* (2024).

[24] Ö. Şahinaslan, H. Dalyan, and E. Şahinaslan., *Multilingual Sentiment Analysis on YouTube Data Using the Naive Bayes Classifier. Journal of Information Technologies*, vol. 15, no. 2, pp. 221–229, 2022, doi:10.17671/gazibtd.999960.

[25] A. McCallum., *A Comparison of Event Models for Naive Bayes Text Classification.* (1998).

[26] V. Jakkula., *Tutorial on Support Vector Machine (SVM).* School of EECS, Washington State University, vol. 37, no. 2.5, p. 3, (2006).

[27] Y. Y. Song and Y. Lu., *Decision Tree Methods: Applications for Classification and Prediction. Shanghai Archives of Psychiatry*, vol. 27, no. 2, pp. 130–135, (2015). doi:10.11919/j.issn.1002-0829.215044

[28] A. Liaw and M. Wiener., *Classification and Regression by randomForest. R News*, vol. 2, no. 3, pp. 18–22, (2002).

[29] F. Y. Osisanwo, J. E. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi., *Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, (2017).

Zeynep Orpek,
KFT Bilişim Sistemleri A.Ş
R&D and Innovation Department
Turkey.
E-mail address: erbasizeynep@gmail.com

and

Adem Orpek
LC Waikiki
Information Technology Department
Turkey.
E-mail address: adem.orpek1@gmail.com

and

Onder Sahinaslan
Maltepe University
Department of Informatics
Turkey.
E-mail address: ondersahinaslan@maltepe.edu.tr