COMMENTCLASSIFYR

Senior Design Project II

Tuana Selen Özhazday - Yiğit Can Çelik - Bayram Yağcı

2024

MEF UNIVERSITY
FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

COMMENTCLASSIFYR

Senior Design Project II

Tuana Selen Özhazday - Yiğit Can Çelik - Bayram Yağcı

Advisor: Tuna Çakar, Ph.D.

2024

MEF UNIVERSITY
FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

Project Title : CommentClassifyR
Student(s) Name : Tuana Selen Özhazday - Yiğit Can Çelik - Bayram Yağcı
Date : 24/04/2024

I hereby state that the design project prepared by Tuana Selen Özhazday - Yiğit Can Çelik - Bayram Yağcı has been completed under my supervision. I accept this work as a "Senior Design Project".

24/04/2024
Tuna Çakar, Ph.D.

I hereby state that I have examined this senior design project by Tuana Selen Özhazday - Yiğit Can Çelik - Bayram Yağcı. This work is acceptable as a "Senior Design Project".

24/04/2024
Department Chair's Name

Head of the Department of
Computer Engineering

ACADEMIC HONESTY PLEDGE

In keeping with the MEF University Student Code of Conduct, I pledge that this work is my own and that I have not received inappropriate assistance in its preparation. I further declare that all resources are explicitly cited.

NAME	DATE	SIGNATURE
Tuana Selen Özhazday	24/04/2024	
Yiğit Can Çelik	24/04/2024	
Bayram Yağcı	24/04/2024	

ABSTRACT

COMMENTCLASSIFYR

Tuana Selen Özhazday - Yiğit Can Çelik - Bayram Yağcı

MEF UNIVERSITY
Faculty of Engineering
Department of Computer Engineering

Advisor: Tuna Çakar, Ph.D.

JANUARY 2024

This project aims to facilitate the categorization of product reviews using UiPath automation and machine learning techniques. Initially, the program will leverage UiPath's web scraping capabilities to ethically collect product reviews from the Amazon platform. Subsequently, the program will develop a machine-learning model using a dataset containing 27,000 categorized reviews from Trendyol. These categorized reviews from Trendyol have been obtained using UiPath. The model will be trained using various methods. Once the best approach is determined, this trained model will categorize the newly collected Amazon reviews according to the learned categories, utilizing UiPath's Python integration. The categorized reviews will be compiled into an Excel report, with each category occupying a separate sheet, facilitated by UiPath's Excel integration. A user-friendly Graphical User Interface (GUI) will also be developed to enable users to select and compare different machine learning models for the categorization process.

Keywords: UiPath, Robotic Process Automation (RPA), Machine Learning, Categorization, Web Scraping

ÖZET

COMMENTCLASSIFYR

Tuana Selen Özhazday - Yiğit Can Çelik - Bayram Yağcı

MEF ÜNİVERSİTESİ
Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Tuna Çakar, Ph.D.

OCAK, 2024

Bu proje, UiPath otomasyonu ve makine öğrenimi teknikleri kullanarak ürün incelemelerinin kategorize edilmesini kolaylaştırmayı amaçlamaktadır. İlk olarak, UiPath'ın web kazıma yeteneklerinden yararlanarak, program Amazon platformundan ürün incelemelerini etik bir şekilde toplayacaktır. Daha sonra, Trendyol'dan 27.000 kategorize inceleme içeren bir veri seti kullanarak, program bir makine öğrenimi modeli geliştirecek. Bu kategorize Trendyol incelemeleri, UiPath kullanılarak elde edilmiştir. Model, çeşitli farklı yöntemler kullanılarak eğitilecek. En iyi yöntem belirlendikten sonra, UiPath'in Python entegrasyonu kullanılarak, bu eğitilmiş model, yeni toplanan Amazon incelemelerini öğrenilen kategorilere göre kategorize edecek. Kategorize edilen incelemeler, her kategori için ayrı bir sayfa bulunan Excel raporunda derlenecek, bu da UiPath'in Excel entegrasyonu ile kolaylaştırılacak. Ek olarak, kullanıcıların kategorizasyon süreci için farklı makine öğrenimi modellerini seçmelerini ve karşılaştırmalarını sağlayacak kullanıcı dostu bir Grafik Kullanıcı Arayüzü (GUI) geliştirilecek.

Anahtar Kelimeler: UiPath, Robotik Süreç Otomasyonu, Makine Öğrenimi, Kategori

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GDPR	Data Protection Regulation
CCPA	California Consumer Privacy Act
WBS	Work Breakdown Structure
SVM	Support Vector Machines
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RPA	Robotic Process Automation
UML	Unified Modeling Language

1. INTRODUCTION

In the ever-evolving landscape of digital commerce, the influence of product reviews on consumer behavior and business strategy has become increasingly pronounced. Recognizing this pivotal trend, our project, "CommentClassifyR," embarks on redefining the paradigms of managing and analyzing these critical reviews. At its core, the project ingeniously integrates UiPath's robust automation capabilities with advanced machine learning techniques, aiming to establish a highly efficient, accurate, and streamlined system for categorizing product reviews, particularly from Amazon, a global e-commerce leader.

Our initiative is not merely about leveraging cutting-edge data processing technology; it represents a deeper understanding and strategic utilization of consumer feedback. The project is meticulously structured to commence with ethical and efficient data collection through UiPath's advanced web scraping tools, ensuring adherence to Amazon's usage policies. The crux of our endeavor is developing a robust machine-learning model built upon a foundation of 27,000 categorized reviews from Trendyol. This model, refined through various training methodologies, is crucial in deciphering complex consumer emotions and preferences embedded within studies. Upon identifying the most effective training method, this model will be adeptly applied to categorize new Amazon reviews precisely, facilitated by UiPath's Python integration. The final output, a detailed Excel report, will categorize these reviews into distinct segments, each meticulously organized in its dedicated sheet. Beyond data processing, our project aspires to transcend to a more user-interactive dimension. This feature underscores our commitment to making sophisticated data analysis accessible to a broader audience, including those with limited technical expertise.

In essence, "CommentClassifyR" is poised to be a pivotal tool in e-commerce, significantly enhancing businesses' strategic planning and customer engagement efforts in the digital era. It is a testament to the fusion of technological innovation and a user-centered approach to reshaping the landscape of consumer analysis and business intelligence.

1.1. Motivation

In a landscape where online reviews significantly influence consumer behavior and business performance, quickly and accurately categorizing these reviews is not just a convenience but a necessity. This project addresses a critical gap in market analytics strategies by automating and refining the review categorization process. Businesses, ranging from small enterprises to large corporations, stand to gain immensely from the insights derived from categorized reviews. These insights can guide product development, marketing strategies, customer service improvements, and overall business planning. For consumers, the benefits are equally significant. By simplifying and streamlining the process of review categorization, our solution empowers businesses to respond more effectively to customer needs and preferences, fostering a more dynamic and responsive marketplace.

1.2. Broad Impact

Year	Method	Dataset
2023	Web Scraping and Data Collection	Amazon Product Reviews (randomly selected each time by admin)
2023	Machine Learning Model Development	Trendyol Reviews (almost 27,000 categorized reviews)

Table 1. Overview of Project Methodologies and Datasets

1.2.1. Global Impact of the Solution

The global impact of this project is significant, particularly in promoting sustainable business practices and environmental consciousness. By automating the process of collecting and categorizing product reviews, it reduces the ecological footprint traditionally associated with manual data processing. This efficiency enables businesses worldwide to quickly adapt to consumer feedback, potentially leading to more environmentally friendly products and practices. Overall, this project not only streamlines a vital business function but also supports the broader goal of sustainability in the digital marketplace.

1.2.2. Economic Impact of the Solution

The economic impact of this project is substantial, primarily in terms of enhancing business efficiency and decision-making. Automating the categorization of product reviews significantly reduces the time and labor costs associated with manual data analysis. This efficiency can lead to faster response times to market trends and customer feedback, allowing businesses to adapt more quickly to consumer needs and preferences. Improved product alignment with customer expectations can increase sales and customer loyalty. Additionally, the ability to accurately analyze large data sets with minimal resource expenditure makes this technology a cost-effective solution for businesses of all sizes. Thus, this project not only boosts operational efficiency but also has the potential to drive economic growth and innovation across various sectors.

1.2.3. Environmental Impact of the Solution

The project positively impacts the environment by promoting digital automation and reducing reliance on resource-intensive processes. By enabling remote and efficient data processing, it decreases the environmental footprint associated with traditional office-based work, such as energy usage and commuting-related emissions. Additionally, the insights gained from consumer feedback can guide businesses towards more environmentally friendly products and practices, encouraging a market shift towards sustainability. Overall, the project aligns with eco-friendly initiatives, reducing resource usage and supporting a more sustainable approach to business operations.

1.2.4. Societal Impacts of the Solution

The societal impacts of this project are profound, particularly in enhancing consumer empowerment and business transparency. By efficiently categorizing and analyzing product reviews, the project provides consumers with more accurate and easily accessible information, enabling informed purchasing decisions. This transparency fosters trust between consumers and businesses, strengthening the consumer-business relationship. In essence, this project improves business operations and plays a pivotal role in shaping a more informed consumer society.

1.2.5. Legal Issues Related to the Project

In developing this project, a conscientious appraisal of legal and ethical considerations is paramount, particularly concerning health, security, and business practices. Foremost among these is adherence to data privacy and security laws, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). The project ensures the ethical collection, storage, and processing of user data, aligning with legal mandates that require user consent for data usage and upholding stringent data security standards to prevent breaches. Intellectual property rights are another critical area of legal concern. The project navigates the complexities of using publicly available data, such as product reviews, while respecting the copyrights that may be held by the authors or the platforms hosting these reviews.

The project also upholds a strong commitment to human rights, ensuring that its applications do not discriminate or harm individuals or groups. This commitment extends to ensuring that the technology is employed in a manner that respects individual rights and promotes equitable treatment. Finally, business compliance is rigorously observed. The project adheres to relevant business laws and e-commerce regulations, ensuring accurate and ethical data representation and avoiding any form of false advertising or data manipulation based on the analysis conducted.

2. PROJECT DEFINITION AND PLANNING

2.1. Project Definition

Scope of the Project:

The scope of this thesis project encompasses developing and implementing an integrated system that automates the categorization of product reviews utilizing UiPath automation and advanced machine learning techniques.

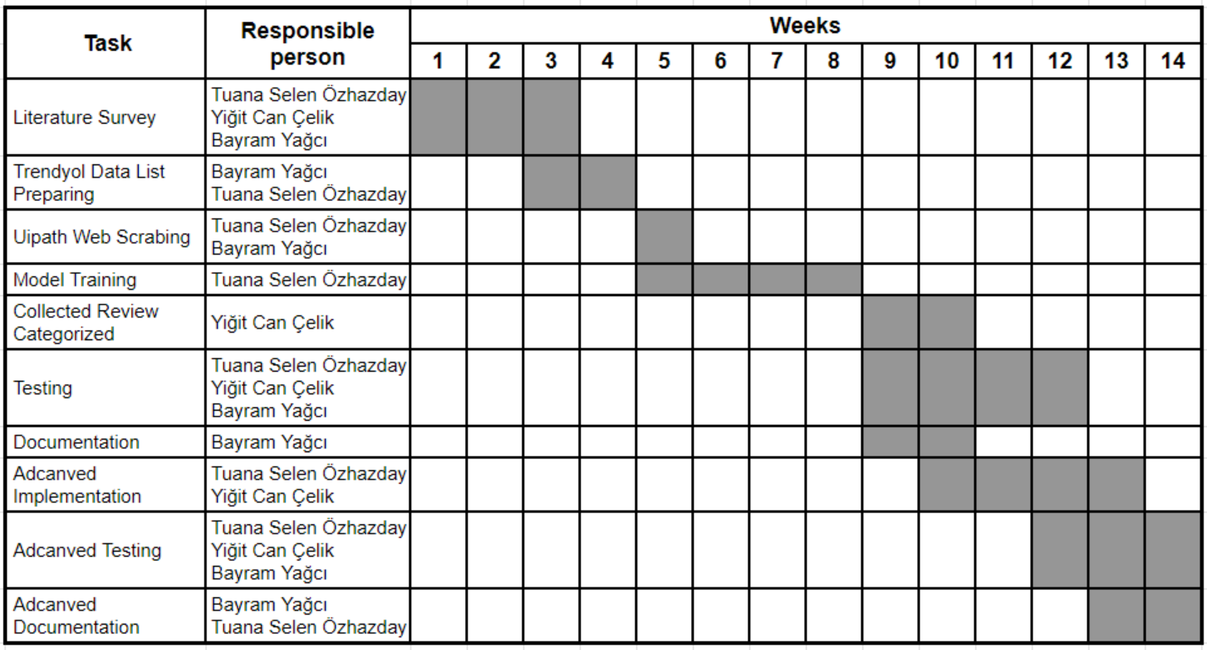
Functional Requirements:

- The system must proficiently scrape and accumulate product reviews from Amazon.
- The developed machine learning model should categorize reviews with high accuracy and minimal error rate.
- Effective integration with the UiPath Python environment is essential for model application.
- The system should automatically generate an Excel report with clear and concise categorization of reviews.

Non-Functional Requirements:

- The system should accommodate an expanding volume of data over time. (Scalability)
- The system must maintain high-performance levels, particularly in processing large datasets. (Optimal Performance)
- Robust security measures are required to safeguard sensitive information collected during the data scraping. (Data Security)
- The system should exhibit consistent performance with minimal downtime. (System Reliability)

2.2. Project Planning



Task	Responsible person	Weeks													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Literature Survey	Tuana Selen Özhazday Yiğit Can Çelik Bayram Yağcı	■	■	■											
Trendyol Data List Preparing	Bayram Yağcı Tuana Selen Özhazday			■	■										
Uipath Web Scrabing	Tuana Selen Özhazday Bayram Yağcı					■									
Model Training	Tuana Selen Özhazday					■	■	■	■						
Collected Review Categorized	Yiğit Can Çelik									■	■				
Testing	Tuana Selen Özhazday Yiğit Can Çelik Bayram Yağcı									■	■	■	■		
Documentation	Bayram Yağcı									■	■				
Adcanved Implementation	Tuana Selen Özhazday Yiğit Can Çelik										■	■	■	■	
Adcanved Testing	Tuana Selen Özhazday Yiğit Can Çelik Bayram Yağcı												■	■	■
Adcanved Documentation	Bayram Yağcı Tuana Selen Özhazday													■	■

Table 2. Project Plan For 14 Weeks

2.2.1 Aim of the Project

The project's primary aim is to innovatively harness the capabilities of UiPath automation and advanced machine learning techniques to develop a comprehensive system for efficiently categorizing product reviews. By creating an automated solution that can intelligently categorize thoughts, the project aims to significantly enhance data processing efficiency for businesses, providing them with actionable insights into consumer behavior and preferences. Through this project, we aspire to contribute meaningfully to data science and business intelligence by bridging the gap between vast data resources and practical, actionable insights.

2.2.2 Project Coverage

The project coverage is detailed in several vital areas:

- ***Automated Data Collection:*** The project implements UiPath’s web scraping tools to collect product reviews from Amazon ethically, ensuring adherence to data usage policies and ethical standards.

- ***Machine Learning Model Development:*** Using a dataset of 27,000 categorized reviews from Trendyol, obtained through UiPath, to train and refine the model to accurately categorize studies.

- ***Training Methodologies:*** The project explores various machine learning methods to train the model, focusing on achieving high accuracy and efficiency in categorizing reviews.

- ***Integration of Trained Model:*** Post development, the project integrates the trained machine learning model with UiPath’s Python environment, enabling the categorization of new Amazon reviews.

- ***Data Compilation and Reporting:*** An essential component of the project is the compilation of categorized reviews into an organized Excel report. This process is facilitated by UiPath’s Excel integration features, ensuring that each category of reviews is presented in separate sheets for easy analysis.

- ***Testing and Evaluation:*** The project involves rigorous testing and evaluation phases to ensure the system's functionality, accuracy, and user-friendliness align with the project objectives.

- ***Documentation:*** Comprehensive documentation is provided, detailing the methodologies, processes, and outcomes of the project.

2.2.3 Use Cases

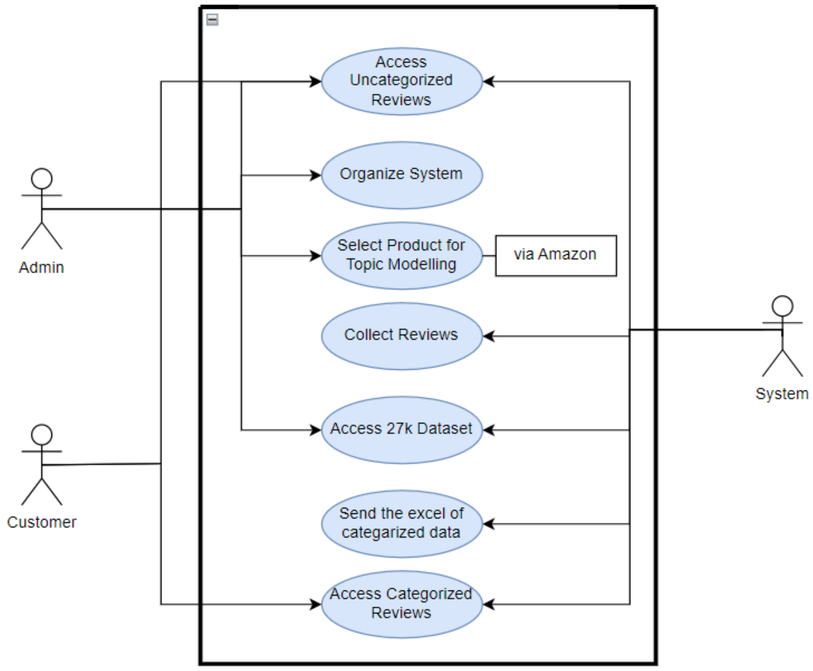

Access Uncategorized Reviews
Organize System
Select Product for Topic Modelling
via Amazon
Collect Reviews
Access 27k Dataset
Send the excel of categarized data
Access Categorized Reviews
Admin
Customer
System

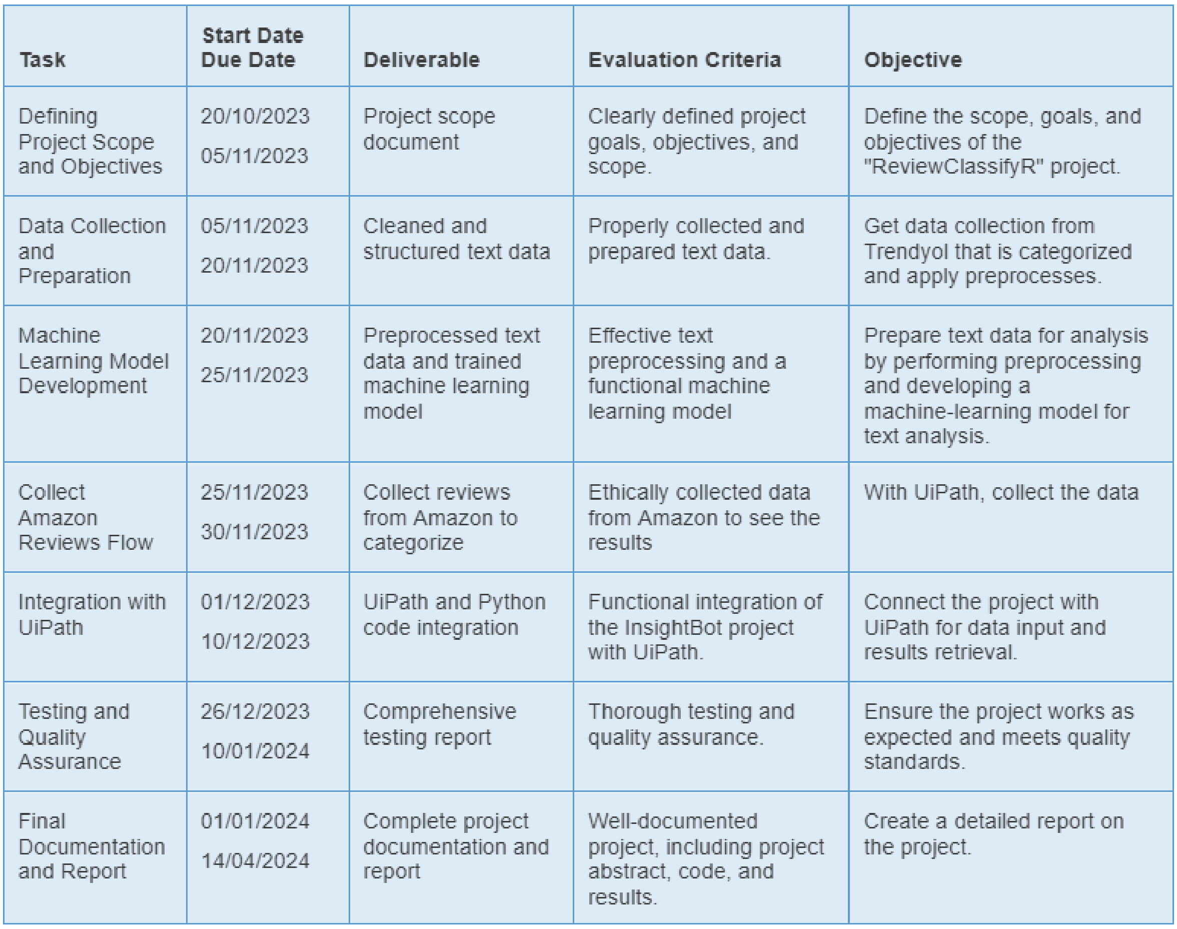
Figure 1. Use Case Diagram

2.2.4 Success Criteria

The project's success hinges on the following key factors:

- The system accurately collects product reviews from Amazon while strictly adhering to ethical standards and legal requirements for data scraping and privacy.
- The accuracy of the machine learning model in categorizing product reviews. High precision and recall rates, indicating fewer misclassifications, will be crucial benchmarks.
- The system's capability to process and categorize large datasets efficiently, maintaining high performance without significant delays or system errors, is a crucial measure of success.
- The project's success will also be evaluated based on its contribution to academic knowledge and practical application in data science and business analytics, including introducing innovative methodologies or insights.

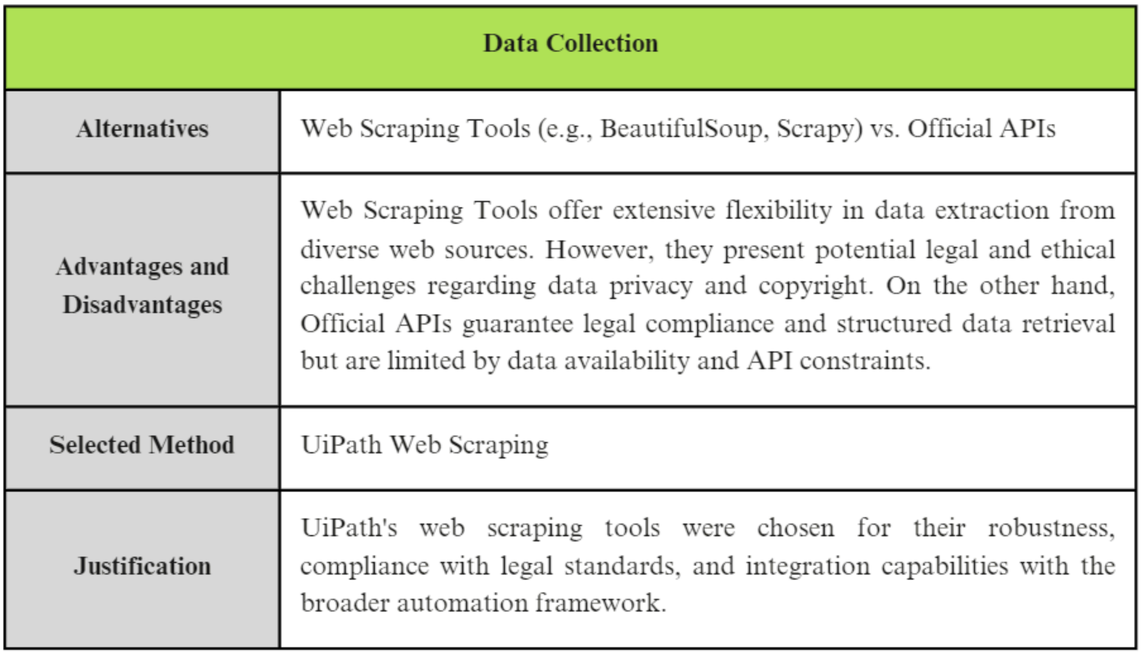
2.2.5 Project Time and Resource Estimation



Task	Start Date Due Date	Deliverable	Evaluation Criteria	Objective
Defining Project Scope and Objectives	20/10/2023 05/11/2023	Project scope document	Clearly defined project goals, objectives, and scope.	Define the scope, goals, and objectives of the "ReviewClassifyR" project.
Data Collection and Preparation	05/11/2023 20/11/2023	Cleaned and structured text data	Properly collected and prepared text data.	Get data collection from Trendyol that is categorized and apply preprocesses.
Machine Learning Model Development	20/11/2023 25/11/2023	Preprocessed text data and trained machine learning model	Effective text preprocessing and a functional machine learning model	Prepare text data for analysis by performing preprocessing and developing a machine-learning model for text analysis.
Collect Amazon Reviews Flow	25/11/2023 30/11/2023	Collect reviews from Amazon to categorize	Ethically collected data from Amazon to see the results	With UiPath, collect the data
Integration with UiPath	01/12/2023 10/12/2023	UiPath and Python code integration	Functional integration of the InsightBot project with UiPath.	Connect the project with UiPath for data input and results retrieval.
Testing and Quality Assurance	26/12/2023 10/01/2024	Comprehensive testing report	Thorough testing and quality assurance.	Ensure the project works as expected and meets quality standards.
Final Documentation and Report	01/01/2024 14/04/2024	Complete project documentation and report	Well-documented project, including project abstract, code, and results.	Create a detailed report on the project.

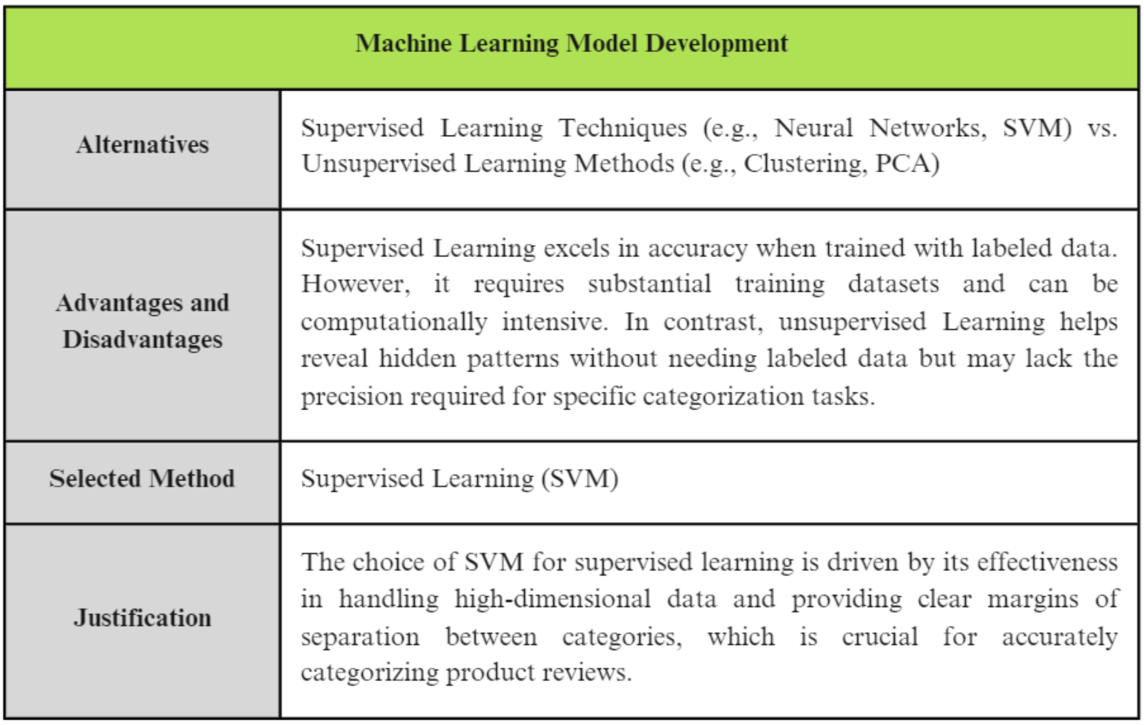
Table 3. The Project’s Time & Resource Estimation

2.2.6 Solution Strategies and Applicable Methods



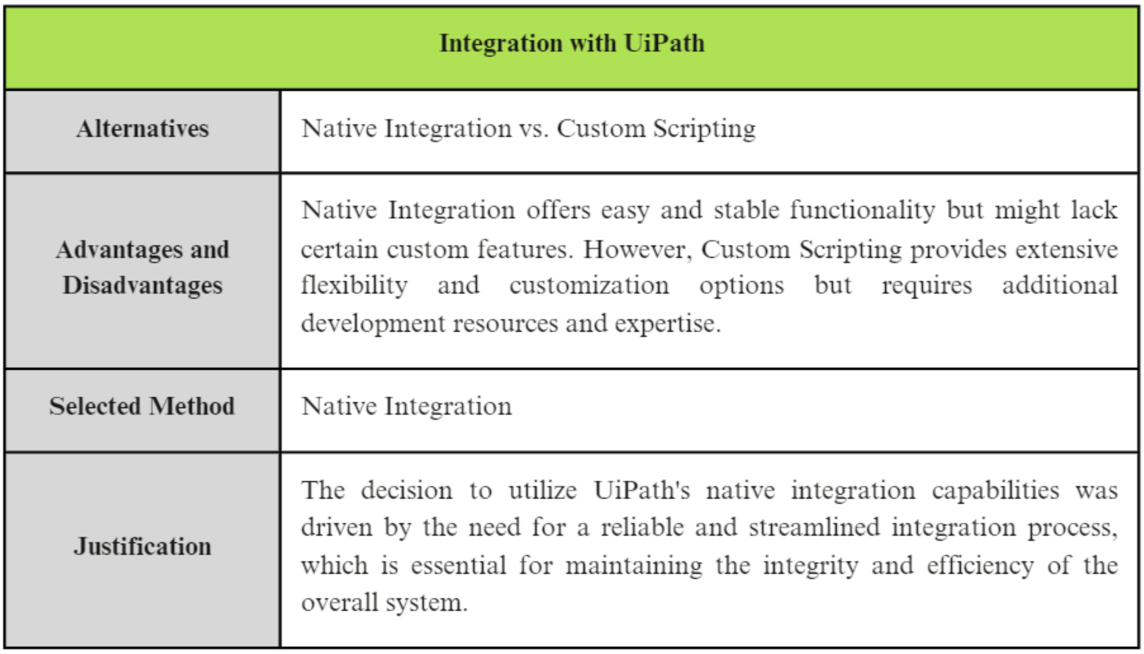
Data Collection	
Alternatives	Web Scraping Tools (e.g., BeautifulSoup, Scrapy) vs. Official APIs
Advantages and Disadvantages	Web Scraping Tools offer extensive flexibility in data extraction from diverse web sources. However, they present potential legal and ethical challenges regarding data privacy and copyright. On the other hand, Official APIs guarantee legal compliance and structured data retrieval but are limited by data availability and API constraints.
Selected Method	UiPath Web Scraping
Justification	UiPath's web scraping tools were chosen for their robustness, compliance with legal standards, and integration capabilities with the broader automation framework.

Table 3. Solution Strategies Table of Data Collection



Machine Learning Model Development	
Alternatives	Supervised Learning Techniques (e.g., Neural Networks, SVM) vs. Unsupervised Learning Methods (e.g., Clustering, PCA)
Advantages and Disadvantages	Supervised Learning excels in accuracy when trained with labeled data. However, it requires substantial training datasets and can be computationally intensive. In contrast, unsupervised Learning helps reveal hidden patterns without needing labeled data but may lack the precision required for specific categorization tasks.
Selected Method	Supervised Learning (SVM)
Justification	The choice of SVM for supervised learning is driven by its effectiveness in handling high-dimensional data and providing clear margins of separation between categories, which is crucial for accurately categorizing product reviews.

Table 4. Solution Strategies Table of Machine Learning Model Development



Integration with UiPath	
Alternatives	Native Integration vs. Custom Scripting
Advantages and Disadvantages	Native Integration offers easy and stable functionality but might lack certain custom features. However, Custom Scripting provides extensive flexibility and customization options but requires additional development resources and expertise.
Selected Method	Native Integration
Justification	The decision to utilize UiPath's native integration capabilities was driven by the need for a reliable and streamlined integration process, which is essential for maintaining the integrity and efficiency of the overall system.

Table 5. Solution Strategies Table of Integration with UiPath

2.2.7 Risk Analysis

1. **Data Privacy and Legal Compliance Violations**

   Impact: High. Non-compliance with data privacy laws, such as GDPR or Amazon's usage policies, can lead to legal repercussions and damage the project's reputation.

   Likelihood: Moderate, given the intricate nature of privacy laws and web scraping.

   Mitigation Strategy: Establish rigorous compliance checks and continuous monitoring of legal developments. Incorporate a legal advisory mechanism into the project structure.

   Contingency Plan: In the event of a compliance issue, immediately halt data collection, consult legal experts, and restructure the collection process in alignment with legal requirements.

2. **Technical Challenges**

   Impact: High, potentially significantly impacting the project's success. Technical challenges may result in project delays or functional limitations.

   Likelihood: Moderate, reasonably likely to occur.

Mitigation Strategy: To address the inefficiency of the Machine Learning Model, Utilize a diversified training dataset and employ continuous model evaluation metrics. To manage Technical Challenges, Conduct extensive pre-deployment testing to identify and rectify issues.

Contingency Plan: Consider alternative integration strategies or tools if technical challenges persist.

2.2.8 Tools Needed to

- Computer: A computer is necessary to process data quickly and for your project to run smoothly.
- UiPath: Automating repetitive tasks speeds up business processes and saves manpower.
- Python: Python is a choice for text mining and machine learning projects and is necessary to perform various tasks at work.
- Machine Learning Libraries: Required for machine learning tasks such as classification and regression. It is aimed at the Scikit-learn library.
- Data Visualization Tools: It is necessary to better understand the data by presenting it visually. Tools such as Matplotlib and Seaborn.
- Machine Learning Models Requirements:
    - Classification Models: Implement and train various classification algorithms suitable for text categorization, such as Naive Bayes, SVM, random forest, or decision trees.
    - Natural Language Processing (NLP) Libraries: Utilize NLP libraries in Python for text preprocessing and feature extraction. NLTK (Natural Language Toolkit) can be used for tokenization, stemming, lemmatization, and stop word removal.

3. THEORETICAL BACKGROUND

3.1. Literature Survey

1. **Model Training with Scikit-learn:**

   In machine learning, Scikit-learn emerges is a pivotal tool for model training, renowned for its flexibility and efficiency. The foundational work of Pedregosa et al. (2011), which presents an extensive exploration of Scikit-learn, demonstrates its applicability in various machine learning tasks, emphasizing its user-friendly interface and wide range of algorithms. Complementing this, the insights provided by James et al. (2013) are crucial for understanding the statistical underpinnings and practical applications of the algorithms within Scikit-learn, particularly in the context of supervised and unsupervised learning techniques. Further, Bass and Bass (2004) delve into generational diffusion models, providing a broader perspective on data behavior and trends essential for informed model training and evaluation. Their work aids in understanding the dynamic nature of data and the predictive modeling capabilities of tools like Scikit-learn.

2. **NLP with NLTK:**

   The NLTK is a cornerstone in NLP, offering a comprehensive suite of libraries and programs for symbolic and statistical natural language processing. The seminal work of Bird, Klein, and Loper (2009) is fundamental in understanding the applications of NLTK, providing detailed guidance on utilizing its tools for a range of NLP tasks. Complementing this, the foundational text by Manning and Schütze (1999) delves into the underlying principles of NLP, offering a comprehensive understanding of the field essential for the effective use of NLTK. Additionally, the work by Jones and Brown (1969) on deep learning techniques sheds light on the processing of large datasets, a capability crucial for NLP tasks handled by NLTK.

3. **Topic Modelling:**

   Topic Modeling is a vital technique in the landscape of NLP, offering profound insights into unstructured textual data. The groundbreaking work on Latent Dirichlet Allocation (LDA) by Blei, Ng, and Jordan (2003) is pivotal in this domain, introducing a probabilistic model that has become foundational in topic modeling. This model provides a methodological framework for understanding and implementing topic extraction algorithms. Complementing this, Griffiths and Steyvers (2004) delve deeper into applying LDA in cognitive science, offering an enhanced perspective on how topic modeling algorithms can reveal hidden thematic structures in large text corpora. Additionally, Bass's (1963) exploration of market share dynamics, while focused on consumer behavior, offers valuable analogies for understanding the dynamic nature of topics in textual data.

4. **UiPath Automation:**

   In Robotic Process Automation (RPA), UiPath is a prominent tool that streamlines complex data workflows and integrates seamlessly with machine learning tasks. Davenport and Ronanki (2018) emphasize the transformative role of RPA in modern business practices, highlighting how platforms like UiPath are revolutionizing data processing and management. Van der Aalst et al. (2018) delve into process mining, illustrating how RPA optimizes and streamlines business processes, a key component of UiPath's functionality. Furthermore, Halevy, Norvig, and Pereira (2009) discuss the intricacies of web data extraction, underlining the significance of UiPath's capabilities in automating these tasks. Zhang et al. (2018) expand on this by exploring the automation of data collection pertinent to understanding UiPath's role in web scraping and data aggregation. Bughin, Hazan, and Ramaswamy (2018) investigate the impact of RPA on enhancing business efficiency and data quality, which aligns with UiPath's goal of providing efficient and accurate data handling solutions. Lastly, Kaplan and Haenlein (2019) address the strategic implications of RPA in decision-making processes relevant to UiPath's contribution to data-driven strategies in various business contexts.

3.2. Automated Review Categorization Using Machine Learning and RPA

In this section, we delineate the solution method adopted for the CommentClassifyR project, which aims to automate the categorization of product reviews. The solution involves a synergistic application of UiPath for RPA, machine learning algorithms for review categorization, and NLP techniques for text data processing.

The following subsections detail the components and workflow of the proposed solution:

1. **Data Collection and Preprocessing:**
    - ***Data Collection with UiPath:*** Utilizing UiPath's web scraping capabilities, we automate the collection of product reviews from the Amazon platform.
    - ***Data Preprocessing:*** The collected data undergoes preprocessing, including text normalization, tokenization, and removal of stop words. This step is crucial for preparing the raw text data for effective machine learning processing.

2. **Machine Learning Model Development:**
    - ***Training Dataset:*** A dataset of 27,000 categorized reviews from Trendyol, obtained through UiPath, forms the basis of our training dataset.
    - ***Model Selection and Training:*** We employ supervised learning algorithms for categorization, particularly Support Vector Machines (SVM).
    - ***Feature Extraction:*** We utilize NLP techniques, leveraging tools such as NLTK and Scikit-learn, for feature extraction from the text data, which includes transforming textual data into a format that is analyzable by the machine learning models.

3. **Integration and Deployment:**
    - ***Integration with UiPath:*** The trained machine learning model is integrated into the UiPath environment. This integration allows the automated categorization of newly collected reviews from Amazon.

4. **Output Generation and Reporting:**
    - ***Categorization Output:*** The system categorizes the reviews based on the trained model, assigning each review to a specific category.
    - ***Reporting:*** The categorized reviews are compiled into an organized Excel report facilitated by UiPath's Excel integration. Each review category is presented separately for easy analysis and interpretation.

4. ANALYSIS AND MODELING

4.1. System Factors

In the CommentClassifyR project, several system factors can significantly influence the system's performance, efficiency, and overall success. Here's an explanation of these critical system factors:

- High-quality, well-structured data ensures more accurate categorization.
- The system must be designed to handle large volumes of data without compromising speed or accuracy.
- The efficiency and accuracy of these algorithms determine the categorization's correctness.
- Scalability is vital to accommodate the growing review data and expand feature requirements.
- Flexibility to adapt to changes in data sources or formats and the ability to integrate new functionalities are crucial for long-term viability.
- The integration of various components, like UiPath for data collection and machine learning models for analysis, must be seamless.
- Compatibility with different platforms and software versions, including Python environments and Excel formats, is essential for smooth operation.
- Adherence to privacy laws and ethical standards, particularly in data collection and storage, is crucial to maintaining user trust and legal compliance.

4.2. How System Works

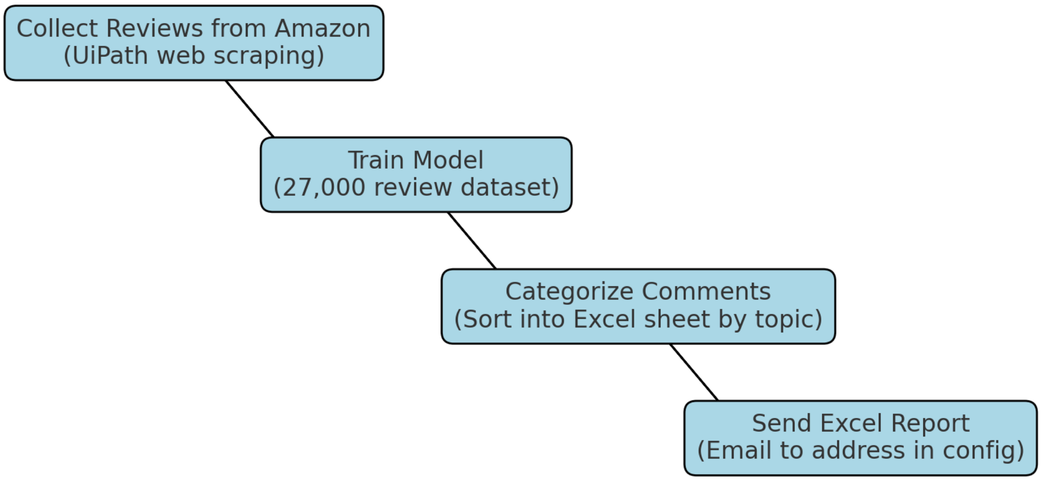
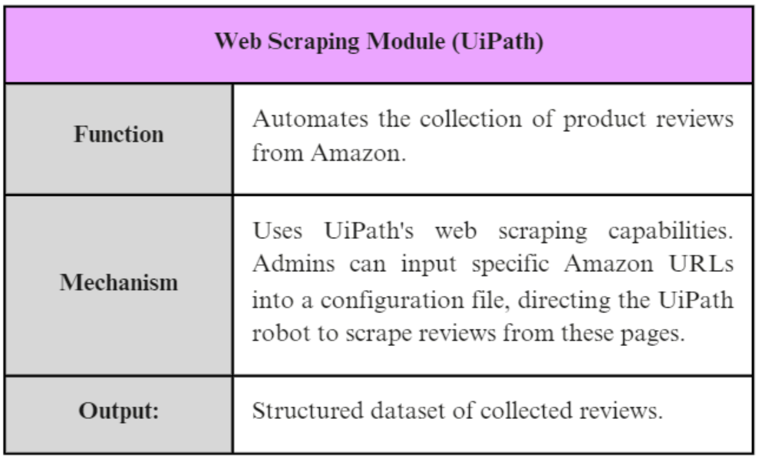


Figure 2. The workflow diagram for the CommentClassifyR system

4.3. Modeling

This model integrates web scraping, machine learning, NLP, and report generation technologies. Here's an explanation of each component within the system model:

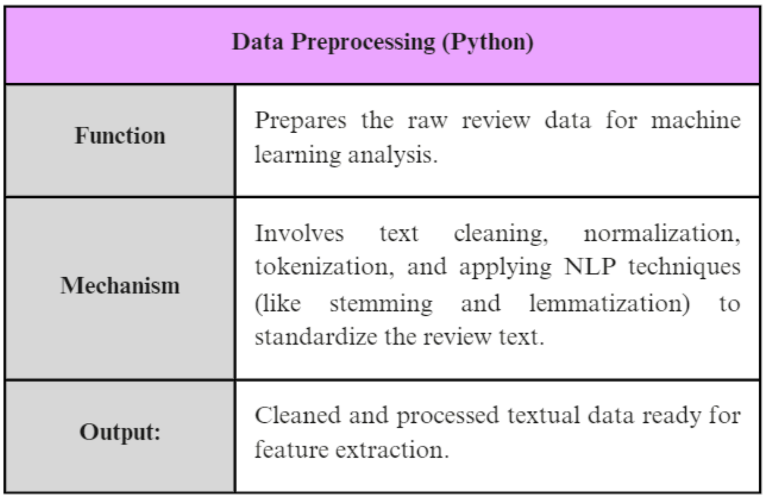
1. ***Data Collection:*** Extracting review data from Amazon, ensuring the data is gathered ethically and in compliance with Amazon's policies.



Web Scraping Module (UiPath)	
Function	Automates the collection of product reviews from Amazon.
Mechanism	Uses UiPath's web scraping capabilities. Admins can input specific Amazon URLs into a configuration file, directing the UiPath robot to scrape reviews from these pages.
Output:	Structured dataset of collected reviews.

Table 6. Modeling Step 1 - Web Scraping Module (UiPath)

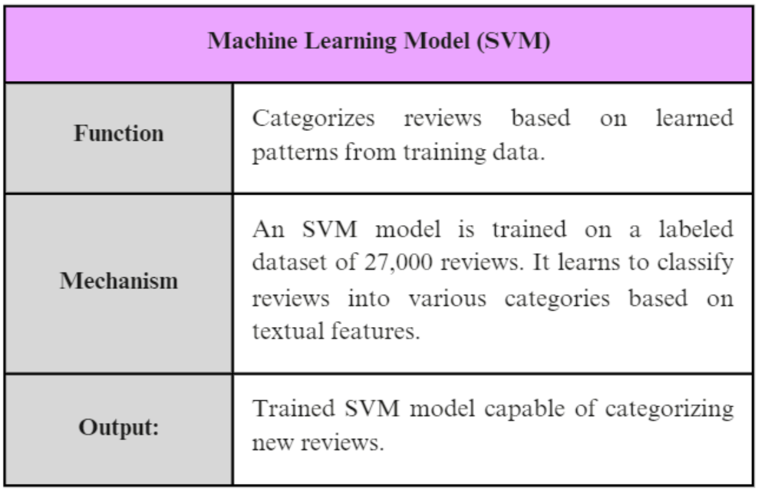
2. ***Data Processing and Preprocessing:*** Cleaning raw data and performing NLP tasks like tokenization and normalization.



Data Preprocessing (Python)	
Function	Prepares the raw review data for machine learning analysis.
Mechanism	Involves text cleaning, normalization, tokenization, and applying NLP techniques (like stemming and lemmatization) to standardize the review text.
Output:	Cleaned and processed textual data ready for feature extraction.

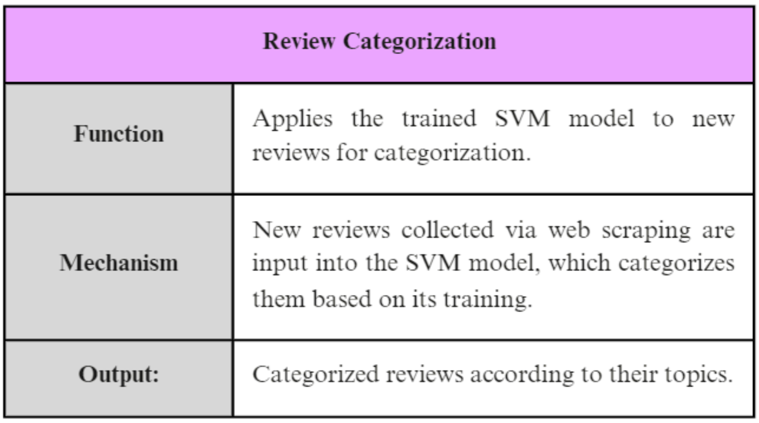
Table 7. Modeling Step 2 - Data Preprocessing (Python)

3. ***Machine Learning Engine:*** Categorizing new reviews based on the trained model, ensuring high accuracy in classification.



Machine Learning Model (SVM)	
Function	Categorizes reviews based on learned patterns from training data.
Mechanism	An SVM model is trained on a labeled dataset of 27,000 reviews. It learns to classify reviews into various categories based on textual features.
Output:	Trained SVM model capable of categorizing new reviews.

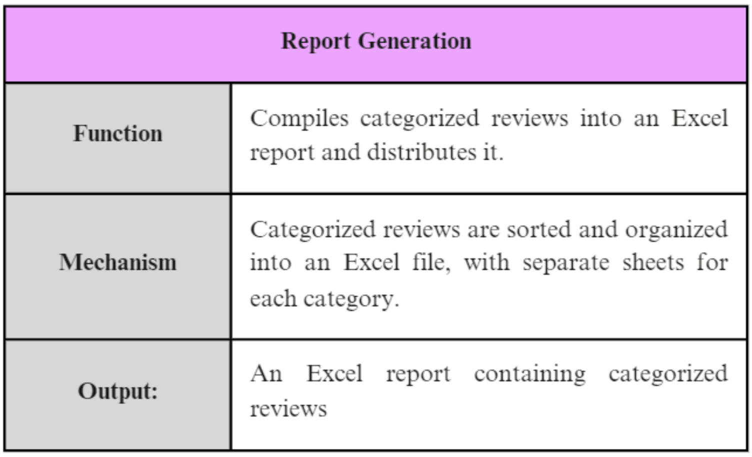
Table 8. Modeling Step 3 - Machine Learning Model (SVM)



Review Categorization	
Function	Applies the trained SVM model to new reviews for categorization.
Mechanism	New reviews collected via web scraping are input into the SVM model, which categorizes them based on its training.
Output:	Categorized reviews according to their topics.

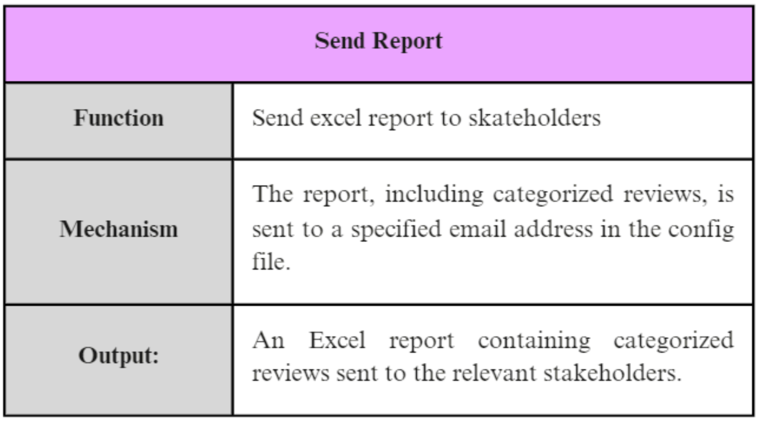
Table 9. Modeling Step 4 - Review Categorization

4. ***Report Generation and Distribution:*** Generating a user-friendly report of the categorized reviews and automating the distribution of this report via email.



Report Generation	
Function	Compiles categorized reviews into an Excel report and distributes it.
Mechanism	Categorized reviews are sorted and organized into an Excel file, with separate sheets for each category.
Output:	An Excel report containing categorized reviews

Table 10. Modeling Step 5 - Report Generation



Send Report	
Function	Send excel report to skateholders
Mechanism	The report, including categorized reviews, is sent to a specified email address in the config file.
Output:	An Excel report containing categorized reviews sent to the relevant stakeholders.

Table 11. Modeling Step 6 - Send Report

4.3.1. System Architecture

The system architecture of the CommentClassifyR project is a structured framework that integrates various components and technologies to automate the process of categorizing product reviews. This architecture ensures efficient data flow, processing, and output generation. The following section is an explanation of the system architecture by UML Diagrams.

4.3.2. UML Diagrams

1. **Use Case Diagram:**

This diagram shows the interactions between administrators, customers, and the system.

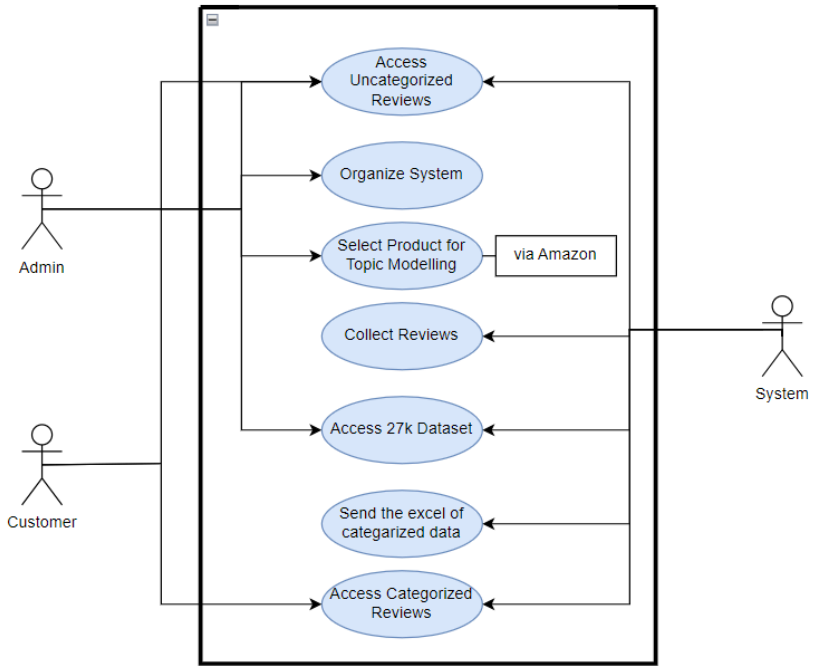

Access Uncategorized Reviews
Organize System
Select Product for Topic Modelling
via Amazon
Collect Reviews
Access 27k Dataset
Send the excel of categarized data
Access Categorized Reviews
Admin
Customer
System

Figure 3. The Use Case Diagram

2. **Class Diagram:**

This diagram details the system's structure in terms of classes, their attributes, methods, and relationships.

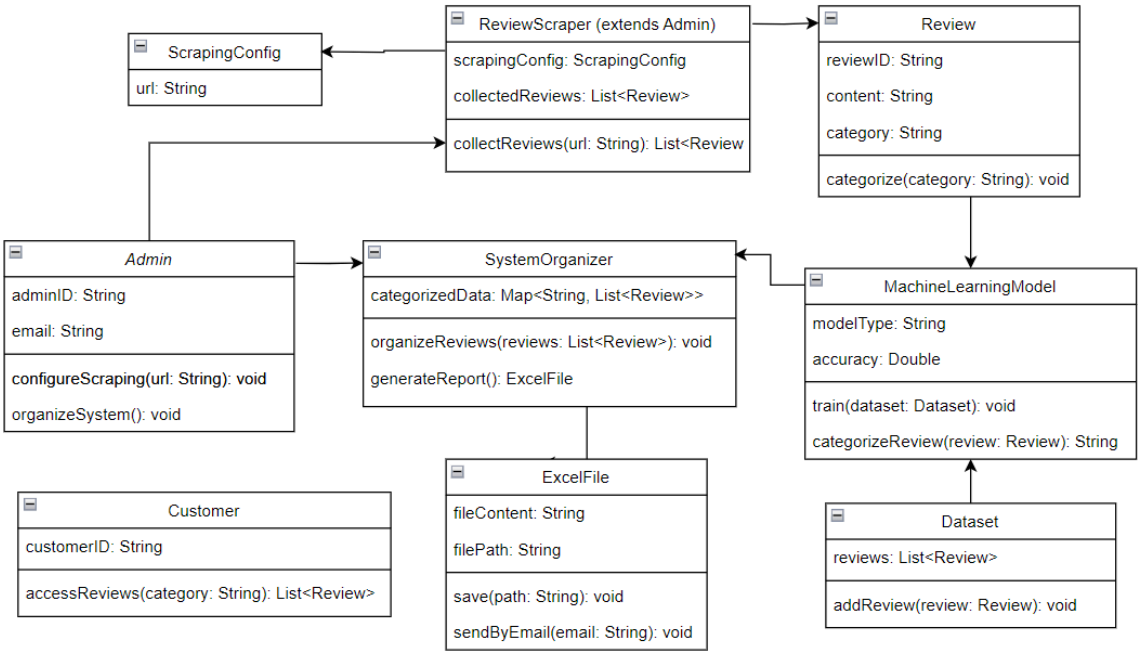

ScrapingConfig
url: String
ReviewScraper (extends Admin)
scrapingConfig: ScrapingConfig
collectedReviews: List<Review>
collectReviews(url: String): List<Review
Review
reviewID: String
content: String
category: String
categorize(category: String): void
Admin
adminID: String
email: String
configureScraping(url: String): void
organizeSystem(): void
SystemOrganizer
categorizedData: Map<String, List<Review>>
organizeReviews(reviews: List<Review>): void
generateReport(): ExcelFile
MachineLearningModel
modelType: String
accuracy: Double
train(dataset: Dataset): void
categorizeReview(review: Review): String
Customer
customerID: String
accessReviews(category: String): List<Review>
ExcelFile
fileContent: String
filePath: String
save(path: String): void
sendByEmail(email: String): void
Dataset
reviews: List<Review>
addReview(review: Review): void

Figure 4. The Class Diagram

3. Sequence Diagram:

This would depict the interaction and the sequence of processes from initiating the review collection to generating reports. It would show objects and the sequence of messages exchanged to carry out the operation.

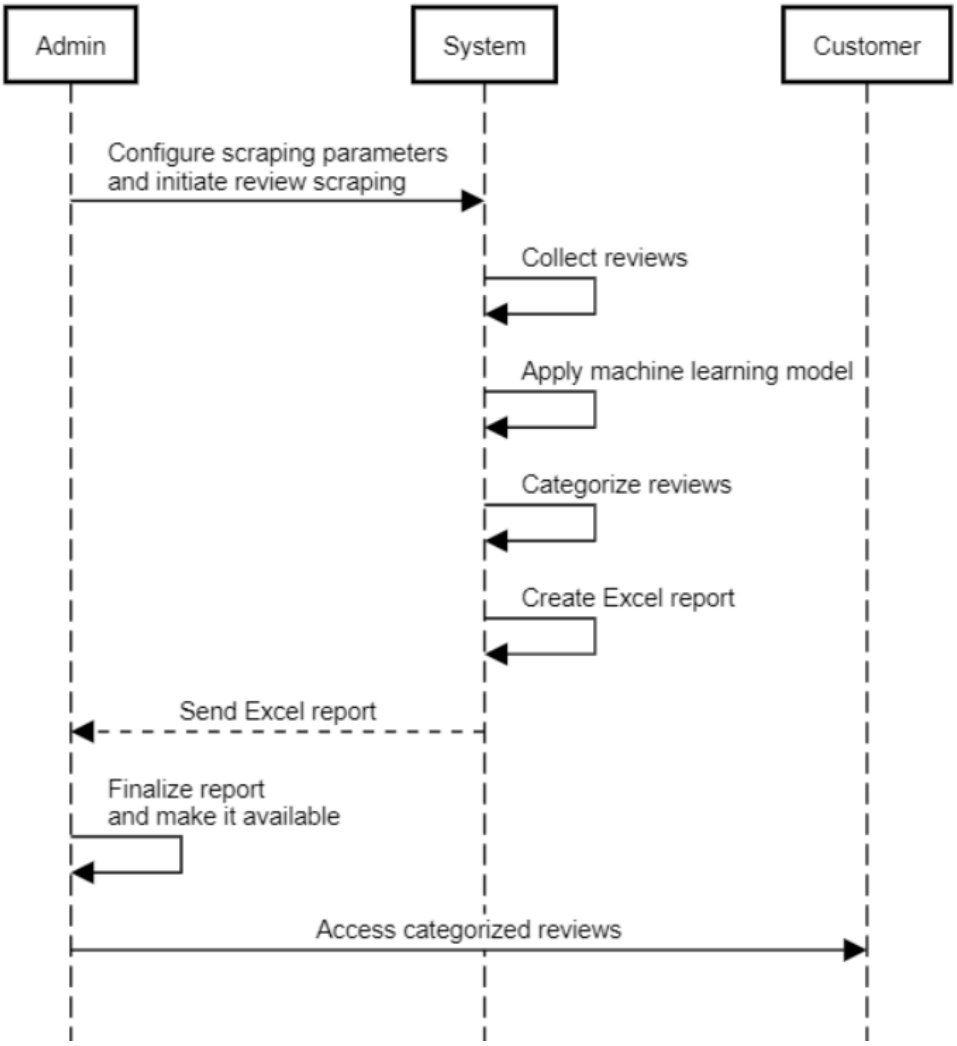

Admin
System
Customer
Configure scraping parameters and initiate review scraping
Collect reviews
Apply machine learning model
Categorize reviews
Create Excel report
Send Excel report
Finalize report and make it available
Access categorized reviews

Figure 5. The Sequence Diagram

4. Activity Diagram:

This diagram would represent the system's workflow, showing the flow from one activity to another. It would illustrate activities like scraping data, processing data, classifying reviews, and generating reports.

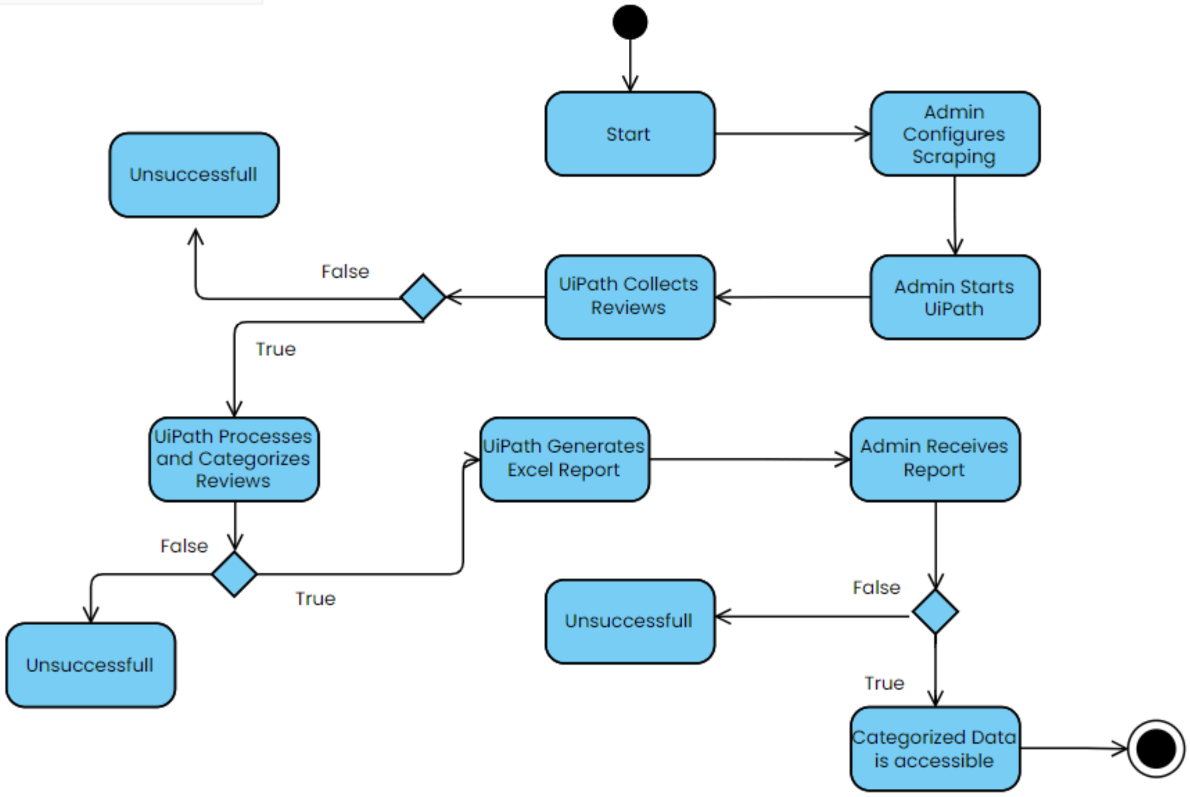


Figure 6. The Activity Diagram

5. DESIGN, IMPLEMENTATION, AND TESTING

5.1 Design

The design of the project includes the following modules:

- ***Data Collection Module:***

The Data Collection module is a fundamental component of the "CommentClassifyR" system. It is responsible for gathering and filtering relevant comment data from various online sources. We utilize UiPath for automating the web scraping process, ensuring efficient and accurate data collection. The module's design focuses on flexibility, allowing for easy adaptation to different data sources. This adaptability is crucial for the robustness of the overall system, as it ensures that our model has access to diverse and comprehensive data sets.

- ***Machine Learning Model Module:***

The crucial of "CommentClassifyR" lies in its Machine Learning Model module. This module is designed to categorize comments based on their content. It employs advanced natural language processing techniques and machine learning algorithms, such as Neural Networks to accurately classify comments. The design of this module emphasizes scalability and performance, allowing for the processing of large datasets while maintaining high accuracy. To achieve this, we have incorporated optimization techniques such as feature selection and hyperparameter tuning.

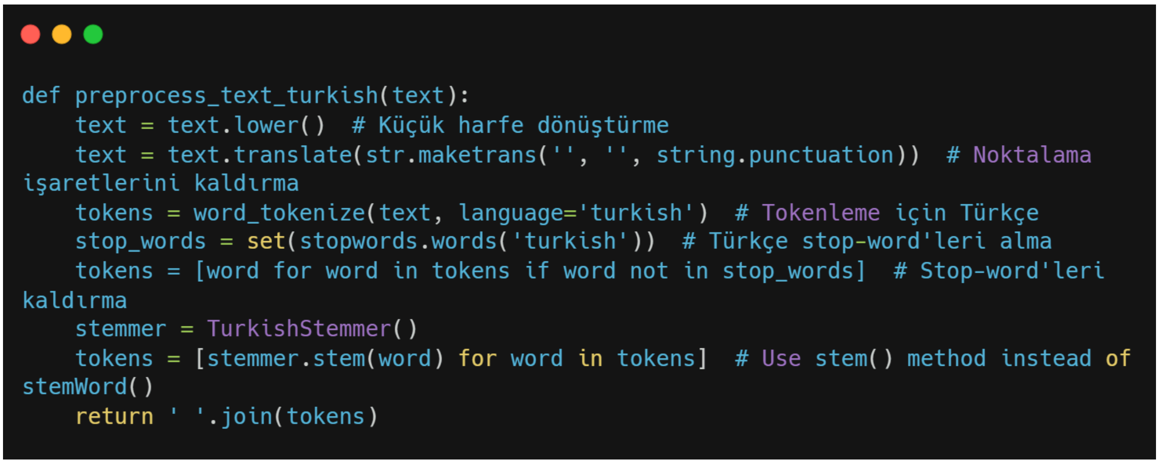
- ***Integration Module:***

The Integration module plays a critical role in ensuring that the Data Collection and Machine Learning Model modules work cohesively. It acts as a bridge, formatting the data collected into a structure that is compatible with the machine learning model. Additionally, it manages the flow of data between these modules, ensuring that each component receives the necessary information in real-time.

5.2 Implementation

5.2.1 Python Seciton

The Python section includes the following key properties:

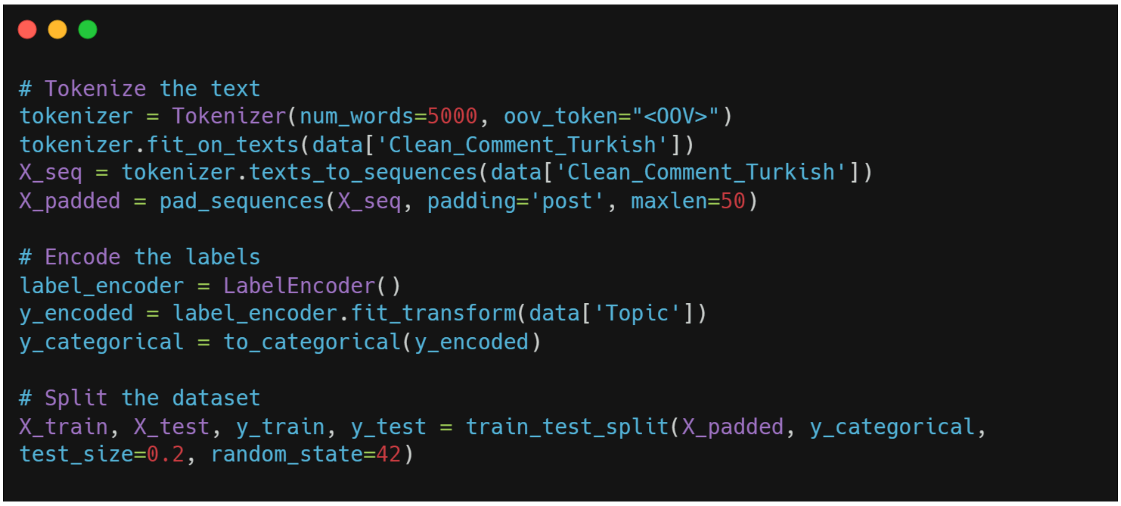


```
def preprocess_text_turkish(text):
    text = text.lower()  # Küçük harfe dönüştürme
    text = text.translate(str.maketrans('', '', string.punctuation))  # Noktalama
işaretlerini kaldırma
    tokens = word_tokenize(text, language='turkish')  # Tokenleme için Türkçe
    stop_words = set(stopwords.words('turkish'))  # Türkçe stop-word'leri alma
    tokens = [word for word in tokens if word not in stop_words]  # Stop-word'leri
kaldırma
    stemmer = TurkishStemmer()
    tokens = [stemmer.stem(word) for word in tokens]  # Use stem() method instead of
stemWord()
    return ' '.join(tokens)
```

Figure 7. Text Preprocessing Function for Turkish Texts

1. Text Preprocessing Function for Turkish Texts = This function is designed to preprocess Turkish text data. It involves several standard steps in natural language processing:

- Converts all characters in the text to lowercase to ensure uniformity.
- Strips out punctuation marks.
- Splits the text into individual words or tokens. Turkish tokenization is applied here.
- Eliminates common Turkish words that don't contribute much to the meaning of the text (like 've' for 'and').
- Reduces words to their base or root form. A Turkish stemmer is used for this purpose.

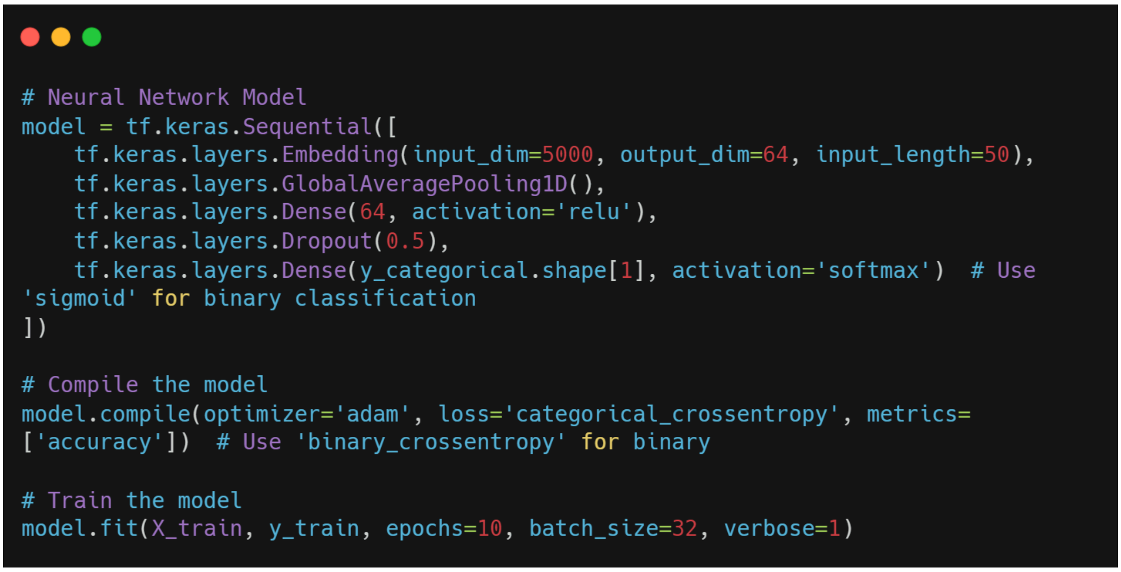


```
# Tokenize the text
tokenizer = Tokenizer(num_words=5000, oov_token="<OOV>")
tokenizer.fit_on_texts(data['Clean_Comment_Turkish'])
X_seq = tokenizer.texts_to_sequences(data['Clean_Comment_Turkish'])
X_padded = pad_sequences(X_seq, padding='post', maxlen=50)

# Encode the labels
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(data['Topic'])
y_categorical = to_categorical(y_encoded)

# Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X_padded, y_categorical,
test_size=0.2, random_state=42)
```

Figure 8. Text Tokenization and Padding, Label Encoding and Data Splitting

2. Text Tokenization and Padding = A Tokenizer is created, which will convert words into numeric tokens. It keeps a maximum of 5000 words. The tokenizer is then fitted on the preprocessed comments. The texts are converted to sequences of tokens. Sequences are padded to ensure that they all have the same length for model input.

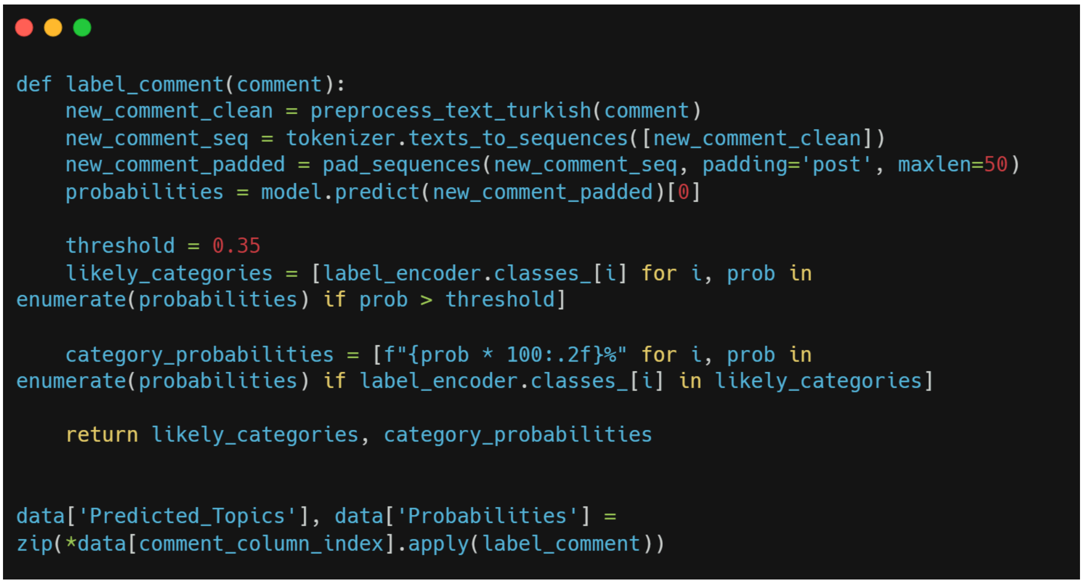
3. Label Encoding and Data Splitting = The topics (labels) in the dataset are encoded into a numerical format using LabelEncoder. These labels are then converted to a one-hot encoded format. The dataset is split into training and testing sets, with 20% of the data reserved for testing.



```
# Neural Network Model
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(input_dim=5000, output_dim=64, input_length=50),
    tf.keras.layers.GlobalAveragePooling1D(),
    tf.keras.layers.Dense(64, activation='relu'),
    tf.keras.layers.Dropout(0.5),
    tf.keras.layers.Dense(y_categorical.shape[1], activation='softmax')  # Use
'sigmoid' for binary classification
])

# Compile the model
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=
['accuracy'])  # Use 'binary_crossentropy' for binary

# Train the model
model.fit(X_train, y_train, epochs=10, batch_size=32, verbose=1)
```

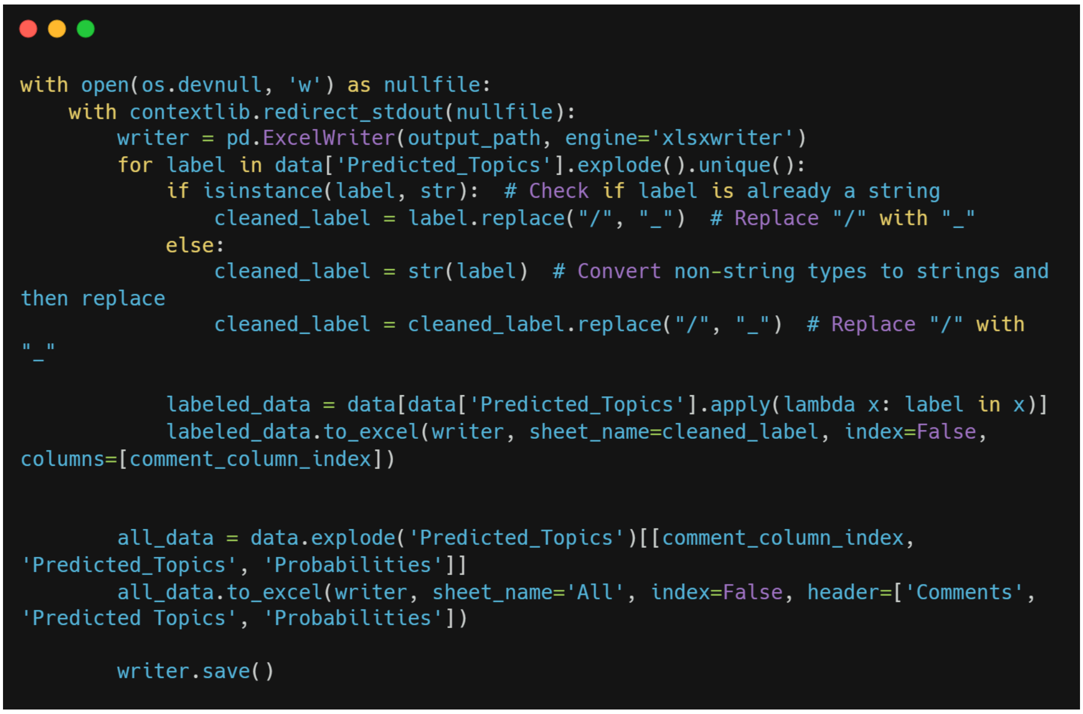
Figure 9. Neural Network Model Definition and Training

4. Neural Network Model Definition and Training = Defines a Sequential neural network model with embedding, global average pooling, dense, dropout layers, and an output layer. The model is compiled with the Adam optimizer and categorical cross-entropy loss function (appropriate for multi-class classification). The model is then trained on the training data.



```
def label_comment(comment):
    new_comment_clean = preprocess_text_turkish(comment)
    new_comment_seq = tokenizer.texts_to_sequences([new_comment_clean])
    new_comment_padded = pad_sequences(new_comment_seq, padding='post', maxlen=50)
    probabilities = model.predict(new_comment_padded)[0]

    threshold = 0.35
    likely_categories = [label_encoder.classes_[i] for i, prob in
enumerate(probabilities) if prob > threshold]

    category_probabilities = [f"{prob * 100:.2f}%" for i, prob in
enumerate(probabilities) if label_encoder.classes_[i] in likely_categories]

    return likely_categories, category_probabilities


data['Predicted_Topics'], data['Probabilities'] =
zip(*data[comment_column_index].apply(label_comment))
```

Figure 10. Labeling New Comments



```
with open(os.devnull, 'w') as nullfile:
    with contextlib.redirect_stdout(nullfile):
        writer = pd.ExcelWriter(output_path, engine='xlsxwriter')
        for label in data['Predicted_Topics'].explode().unique():
            if isinstance(label, str):  # Check if label is already a string
                cleaned_label = label.replace("/", "_")  # Replace "/" with "_"
            else:
                cleaned_label = str(label)  # Convert non-string types to strings and
then replace
                cleaned_label = cleaned_label.replace("/", "_")  # Replace "/" with
"_"

            labeled_data = data[data['Predicted_Topics'].apply(lambda x: label in x)]
            labeled_data.to_excel(writer, sheet_name=cleaned_label, index=False,
columns=[comment_column_index])


        all_data = data.explode('Predicted_Topics')[[comment_column_index,
'Predicted_Topics', 'Probabilities']]
        all_data.to_excel(writer, sheet_name='All', index=False, header=['Comments',
'Predicted Topics', 'Probabilities'])

        writer.save()
```

Figure 11. Exporting New Comments

5. Labeling and Exporting New Comments = This section reads a new dataset of comments from an Excel file. Each comment is then processed, tokenized, and fed into the trained model to predict its topic(s). A threshold of 35% is used to filter the predictions. The predictions and their probabilities are added to the dataset. The modified dataset is written to a new Excel file, with separate sheets for each topic.

5.2.2 Python-Integrated UiPath Section

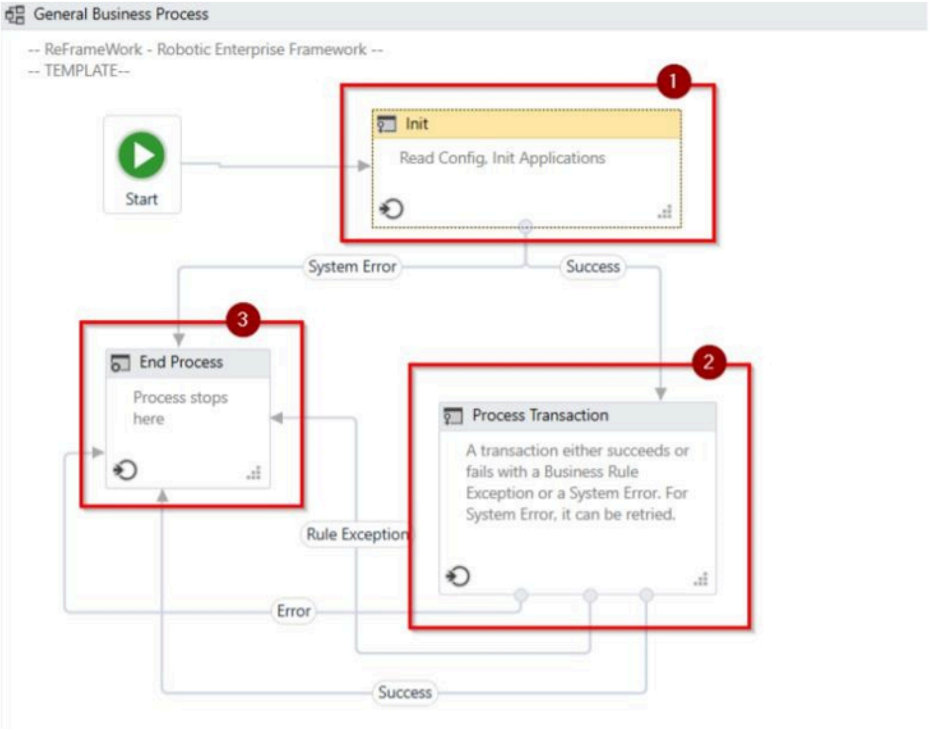
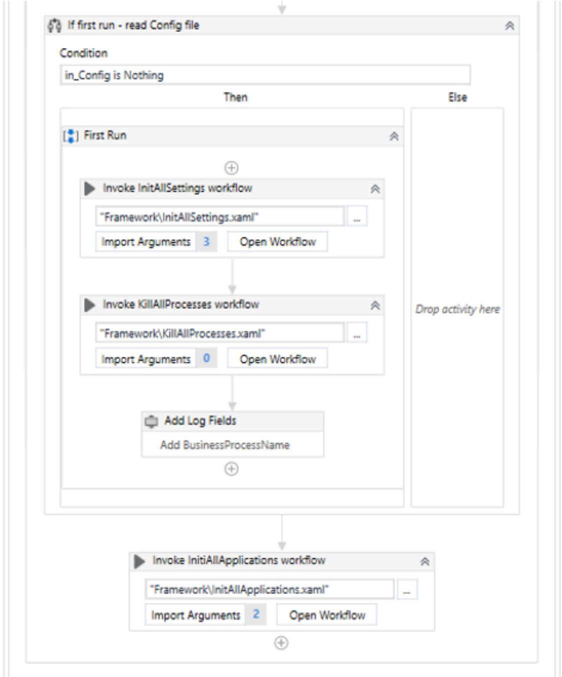

General Business Process
-- ReFrameWork - Robotic Enterprise Framework --
-- TEMPLATE--
1
Init
Read Config, Init Applications
Start
System Error
Success
3
End Process
Process stops here
2
Process Transaction
A transaction either succeeds or fails with a Business Rule Exception or a System Error. For System Error, it can be retried.
Rule Exception
Error
Success

Figure 12. Uipath Main View

Uipath REFramework is used for this project.

- **Init State**



If first run - read Config file

Condition: in_Config is Nothing

Then

First Run

Invoke InitAllSettings workflow: "Framework\InitAllSettings.xaml" — Import Arguments 3

Invoke KillAllProcesses workflow: "Framework\KillAllProcesses.xaml" — Import Arguments 0

Add Log Fields: Add BusinessProcessName

Else

Invoke InitiAllApplications workflow: "Framework\InitAllApplications.xaml" — Import Arguments 2

Figure 13. Uipath Inıt State

The Init state contains all the robot's initial settings.

- **Invoke Init All Settings.xaml**

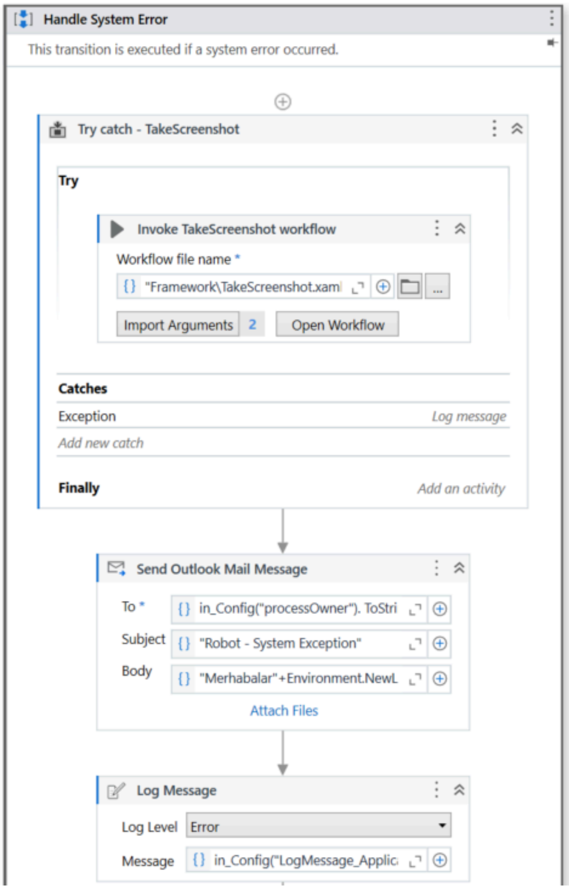
Before the process starts, the config file that has the information about the process is extracted to use when the robot runs.

- **Invoke KillAllProcesses.xaml**

All application exe is killed. In our process, the main applications are SAP and Excel. Therefore, all applications should be killed before the robot starts the process transaction

- **Init Exception Handling**

When the robot is in its initial state, any error that occurs an exception will be captured, taking a screenshot and sending an e-mail process to ProcessOwnerEmail that is defined in config excel.



This transition is executed if a system error occurred.

Workflow file name * "Framework\TakeScreenshot.xaml

Exception: Log message

To * in_Config("processOwner").ToStri

Subject "Robot - System Exception"

Body "Merhabalar"+Environment.NewL

Log Level Error

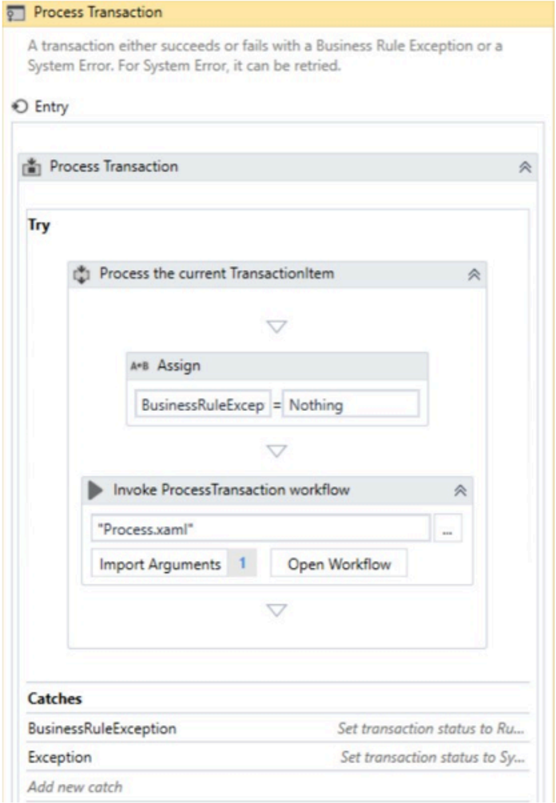
Message in_Config("LogMessage_Applic

Figure 14. Uipath Handle System Error

- ***InitAllApplications workflow***

Before the process starts, the Outlook application should be ready. Firstly, the outlook application is controlled, if it is not running, then the robot will open the outlook.

- **Process Transaction State**


Process Transaction

A transaction either succeeds or fails with a Business Rule Exception or a System Error. For System Error, it can be retried.

Entry

Process Transaction

Try

Process the current TransactionItem

Assign

BusinessRuleExcep = Nothing

Invoke ProcessTransaction workflow

"Process.xaml"

Catches

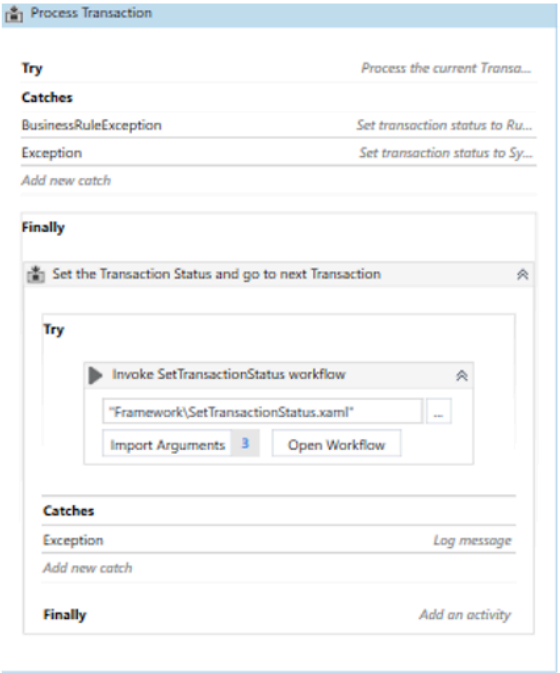
BusinessRuleException Set transaction status to Ru...

Exception Set transaction status to Sy...

Add new catch

Figure 15. Uipath Process State

- ***Exception Handling – Process Transaction***


Process Transaction

Try Process the current Transa...

Catches

BusinessRuleException Set transaction status to Ru...

Exception Set transaction status to Sy...

Add new catch

Finally

Set the Transaction Status and go to next Transaction

Try

Invoke SetTransactionStatus workflow

"Framework\SetTransactionStatus.xaml"

Catches

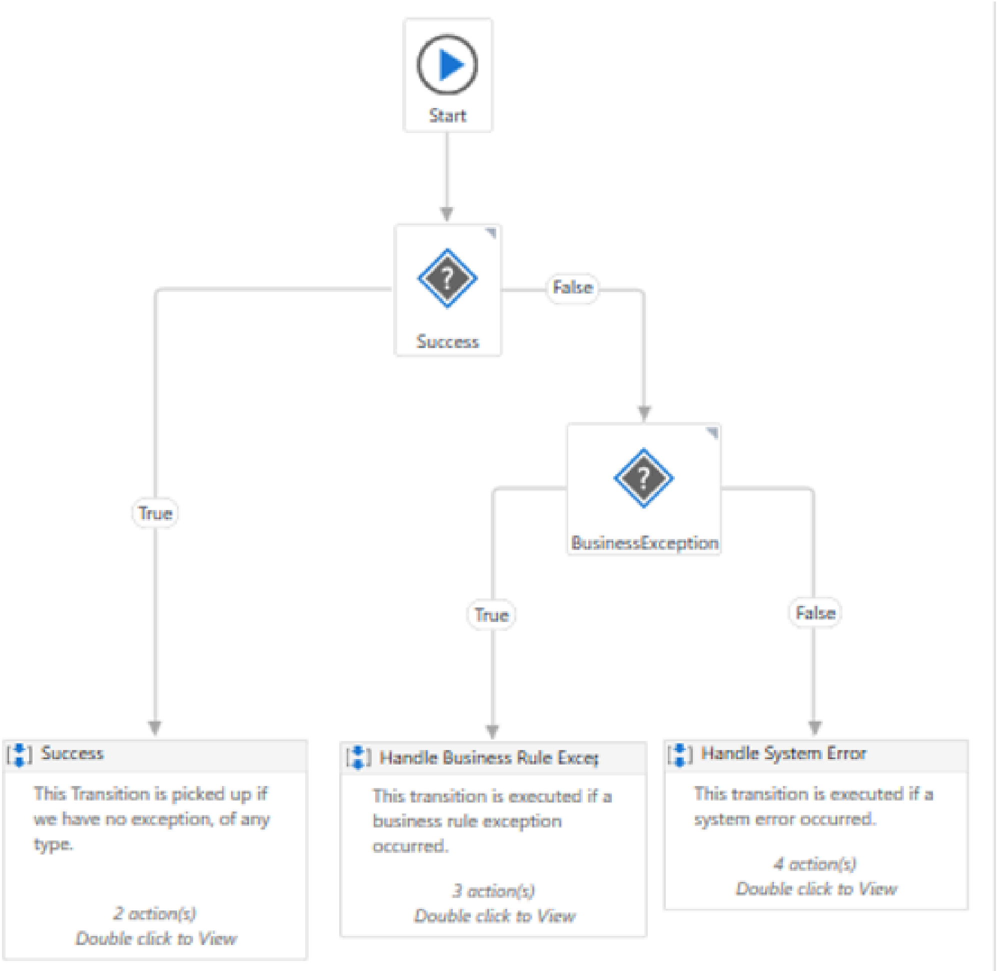
Exception Log message

Add new catch

Finally Add an activity

Figure 16. Uipath Exception Handling

- **Invoke SetTransactionStatus.xaml**



This Transition is picked up if we have no exception, of any type.

2 action(s)
Double click to View

This transition is executed if a business rule exception occurred.

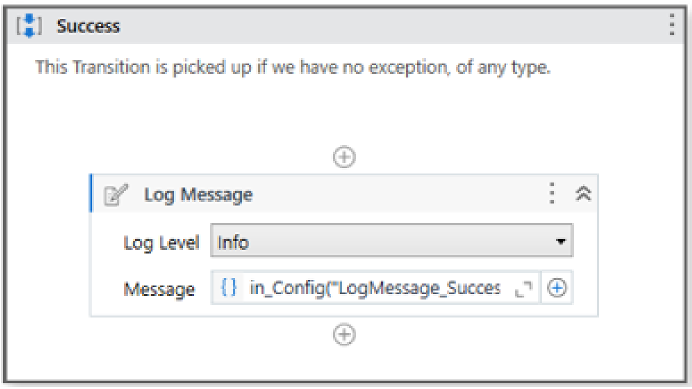
3 action(s)
Double click to View

This transition is executed if a system error occurred.

4 action(s)
Double click to View

Figure 17. Uipath Set Transaction Status

At the end of the process state, if the transaction was finished successfully, then the robot creates a log message.



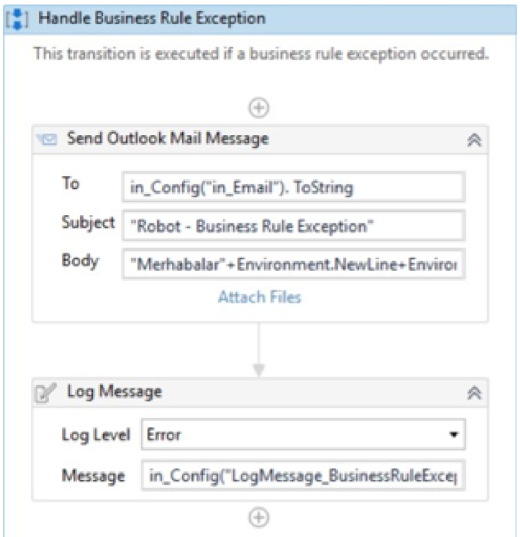
This Transition is picked up if we have no exception, of any type.

Log Level: Info

Message: in_Config("LogMessage_Succes

Figure 18. Uipath Successfully Completed

If there is any error occurs, when the robot is running in the process transaction. Transactions will proceed depending on the exception type. If there is a business exception, then the robot sends an email that contains an exception message.



Handle Business Rule Exception

This transition is executed if a business rule exception occurred.

Send Outlook Mail Message

To: in_Config("in_Email"). ToString

Subject: "Robot - Business Rule Exception"

Body: "Merhabalar"+Environment.NewLine+Enviror

Attach Files

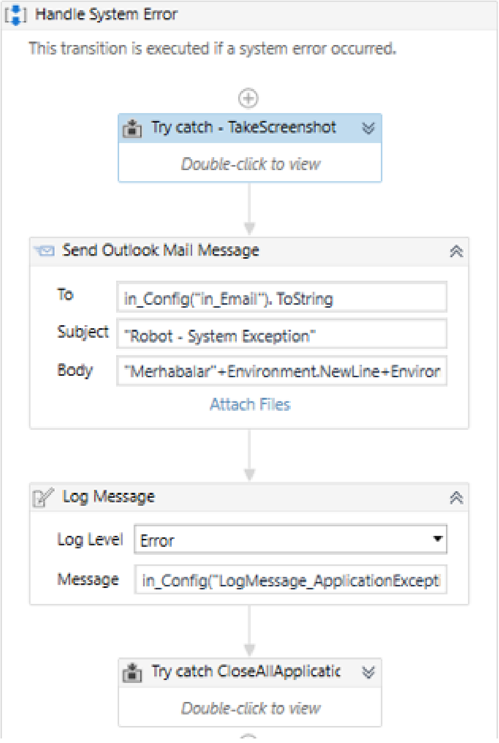
Log Message

Log Level: Error

Message: in_Config("LogMessage_BusinessRuleExce|

Figure 19. Uipath Handle Business Exception

If there is a system exception, then the robot sends an email that contains an exception message.



Handle System Error

This transition is executed if a system error occurred.

Try catch - TakeScreenshot

Double-click to view

Send Outlook Mail Message

To: in_Config("in_Email"). ToString

Subject: "Robot - System Exception"

Body: "Merhabalar"+Environment.NewLine+Enviror

Attach Files

Log Message

Log Level: Error

Message: in_Config("LogMessage_ApplicationExcept

Try catch CloseAllApplicatic

Double-click to view

Figure 20. Uipath Handle System Exception

- **Invoke ProcessTransaction.xaml**

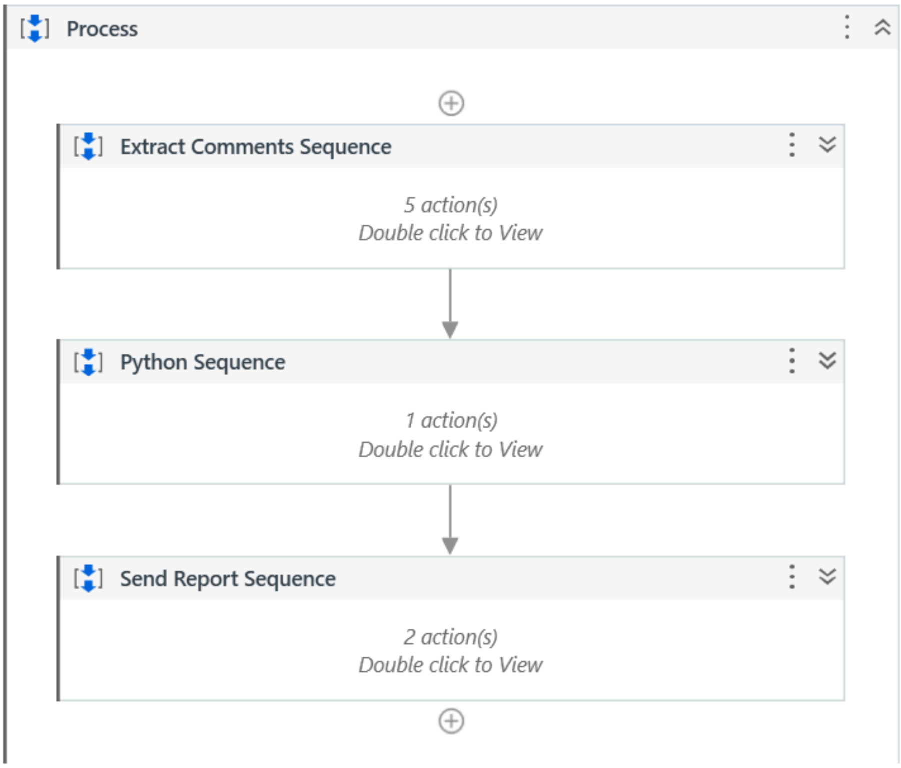
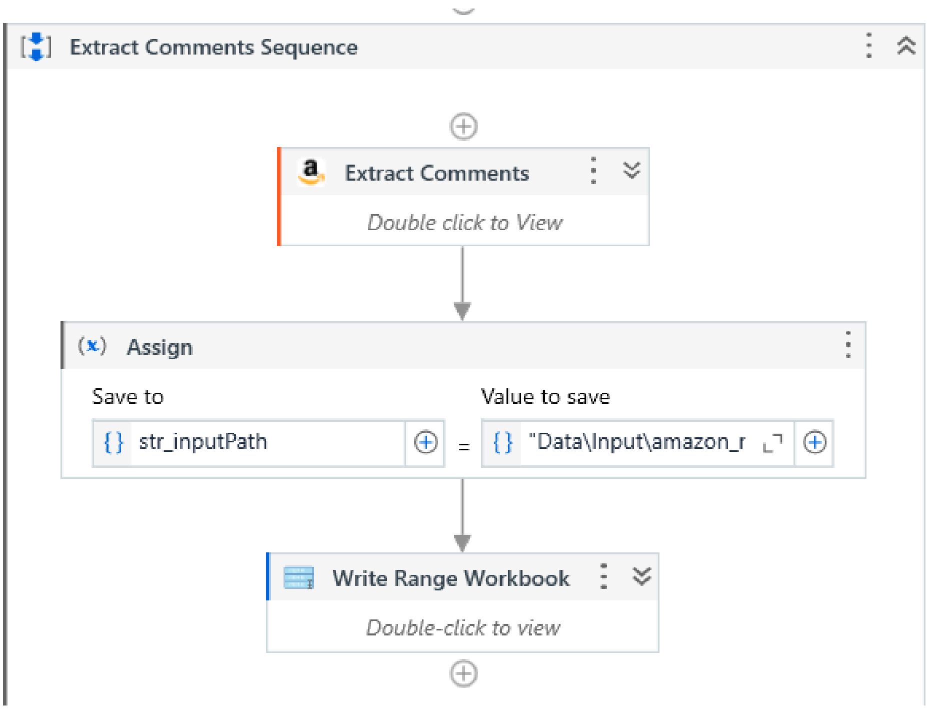
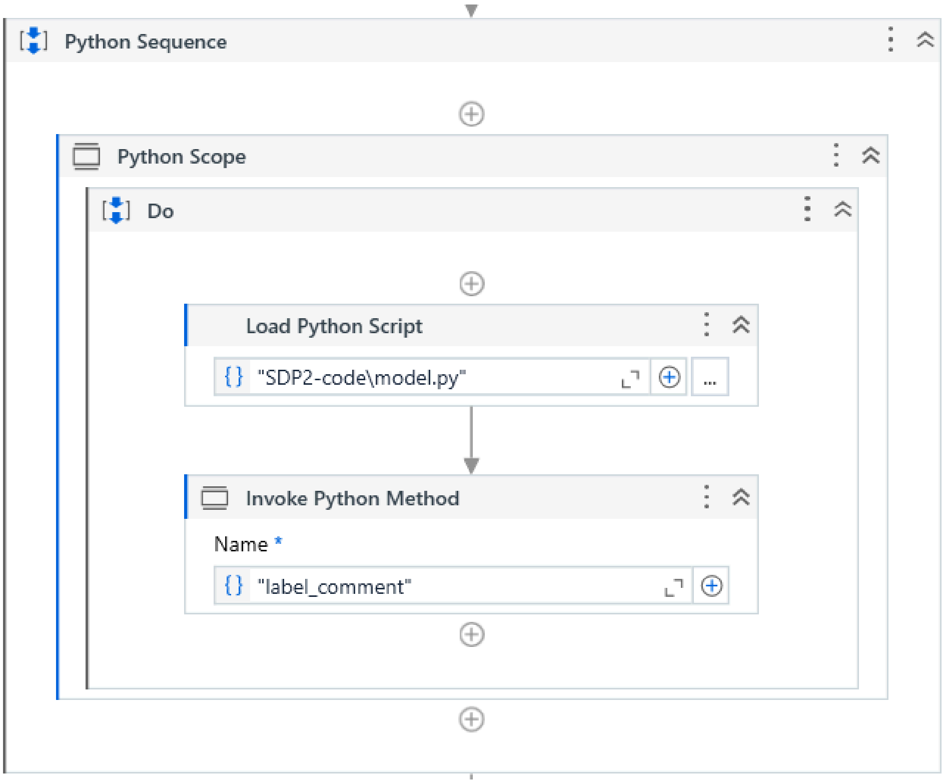


Figure 21. Uipath Process Workflow



Save to: str_inputPath = Value to save: "Data\Input\amazon_r

Figure 22. UiPath Extract Comments Sequence

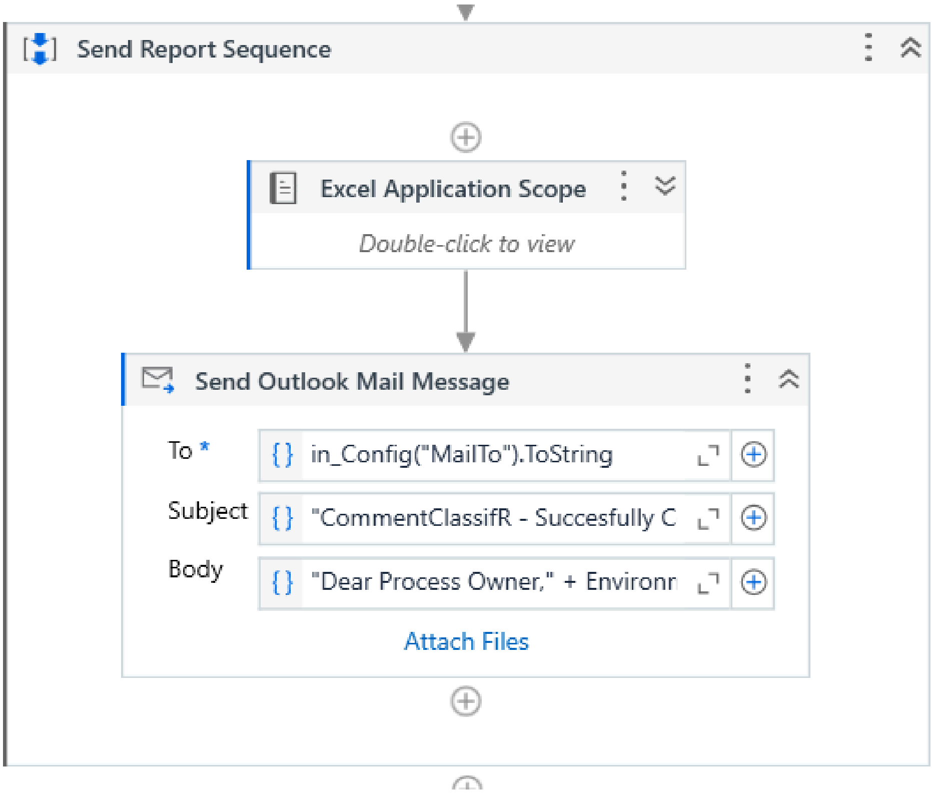


"SDP2-code\model.py"

Name *

"label_comment"

Figure 22. UiPath Python Sequence



To * in_Config("MailTo").ToString

Subject "CommentClassifR - Succesfully C

Body "Dear Process Owner," + Environr

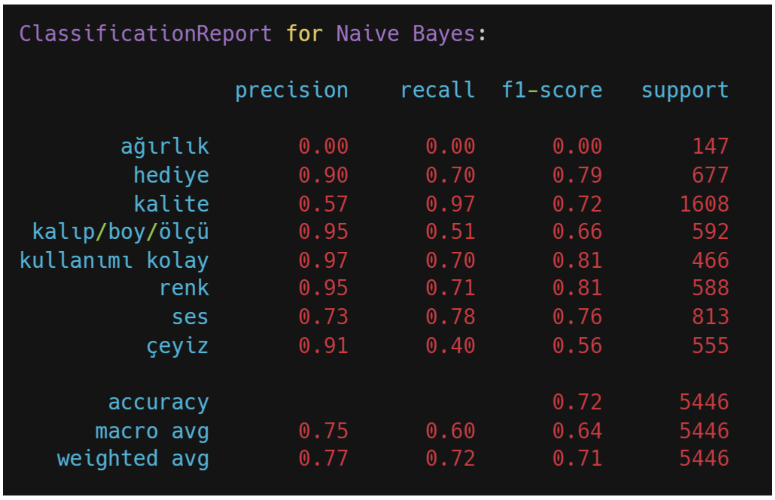
Figure 23. UiPath Send Report Sequence

5.3 Testing

Testing is a critical phase in the creation of machine learning models, when we evaluate the performance and dependability of our models. The primary goal is to assess how well the models generalize to new, untested data. This phase holds paramount importance as it ensures that the models satisfy the anticipated performance criteria and demonstrate robustness suitable for real-world applications.

In our testing, we have employed various standard metrics in the field for performance evaluation. These metrics offer a comprehensive perspective on model efficacy and assist in identifying the most suitable model for deployment. The utilized metrics are as follows:

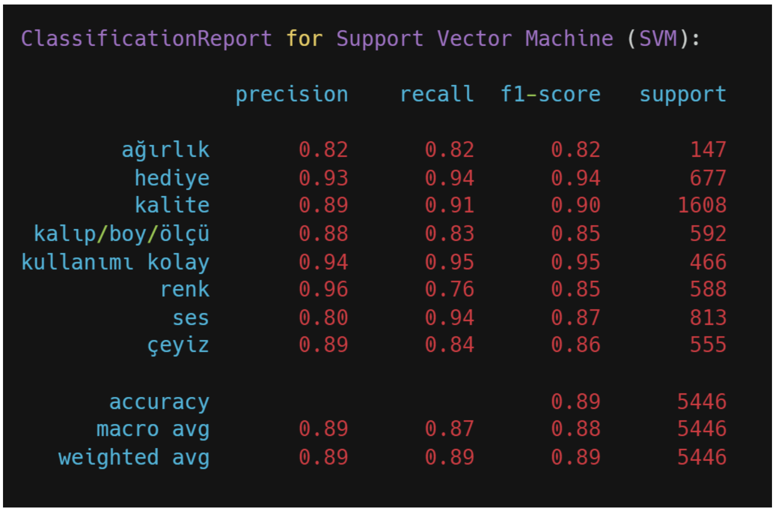
- Precision: This measures how accurate the model's positive predictions were by counting the percentage of true positives among all positive forecasts. Since accuracy denotes a low false positive rate, it is crucial in applications where the cost of a false positive is significant.
- Recall: Also known as sensitivity, this parameter assesses how well the model can recognize all true positives. It is especially crucial in situations where the cost of overlooking a favorable case is substantial.
- F1-Score: The F1-score, which is the harmonic mean of memory and precision, offers a unique statistic that combines recall and precision concerns into a single figure. When type I and type II errors need to be balanced, it is especially helpful.
- Support: The number of true instances for each label in the data is indicated by the support of each class. It is essential for assessing how well the model performs in terms of data distribution.
- Accuracy: This is the percentage of true findings (true positives and true negatives) in relation to all cases that were looked at. It provides a general indicator of the model's efficacy.
- Macro Average: An arithmetic mean of the scores, calculated for each label, and then averaged. It treats all classes equally, irrespective of their frequency in the dataset.
- Weighted Average: This average of the scores is weighted according to the support of each class, considering class imbalance in the dataset.



```
ClassificationReport for Naive Bayes:

                 precision    recall  f1-score   support

        ağırlık       0.00      0.00      0.00       147
         hediye       0.90      0.70      0.79       677
         kalite       0.57      0.97      0.72      1608
 kalıp/boy/ölçü       0.95      0.51      0.66       592
kullanımı kolay       0.97      0.70      0.81       466
           renk       0.95      0.71      0.81       588
            ses       0.73      0.78      0.76       813
          çeyiz       0.91      0.40      0.56       555

       accuracy                           0.72      5446
      macro avg       0.75      0.60      0.64      5446
   weighted avg       0.77      0.72      0.71      5446
```

Table 12. Classification Report for Naive Bayes

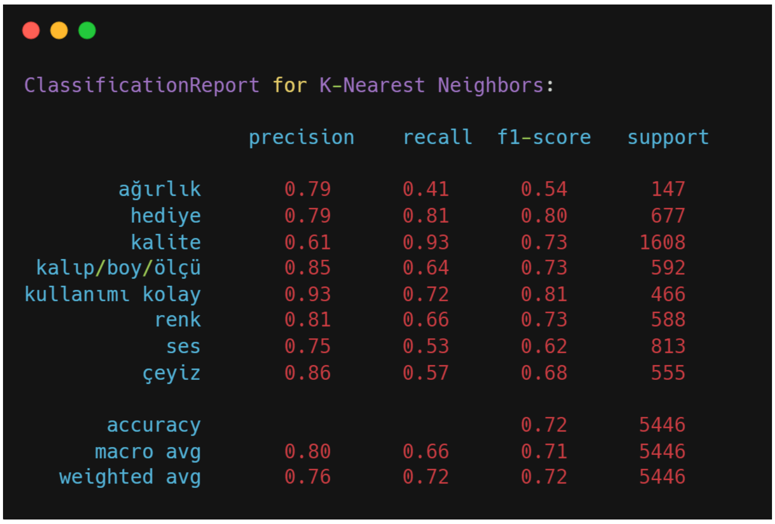
Naive Bayes exhibited significant limitations, most notably in the 'ağırlık' (weight) category with zeros across all metrics, which could be attributed to its assumption of feature independence. This indicates a need for a reevaluation of feature selection or model assumptions, as it may not capture the complexity of the data adequately.



```
ClassificationReport for Support Vector Machine (SVM):

                 precision    recall  f1-score   support

        ağırlık       0.82      0.82      0.82       147
         hediye       0.93      0.94      0.94       677
         kalite       0.89      0.91      0.90      1608
 kalıp/boy/ölçü       0.88      0.83      0.85       592
kullanımı kolay       0.94      0.95      0.95       466
           renk       0.96      0.76      0.85       588
            ses       0.80      0.94      0.87       813
          çeyiz       0.89      0.84      0.86       555

       accuracy                           0.89      5446
      macro avg       0.89      0.87      0.88      5446
   weighted avg       0.89      0.89      0.89      5446
```

Table 13. Classification Report for Support Vector Machine (SVM)

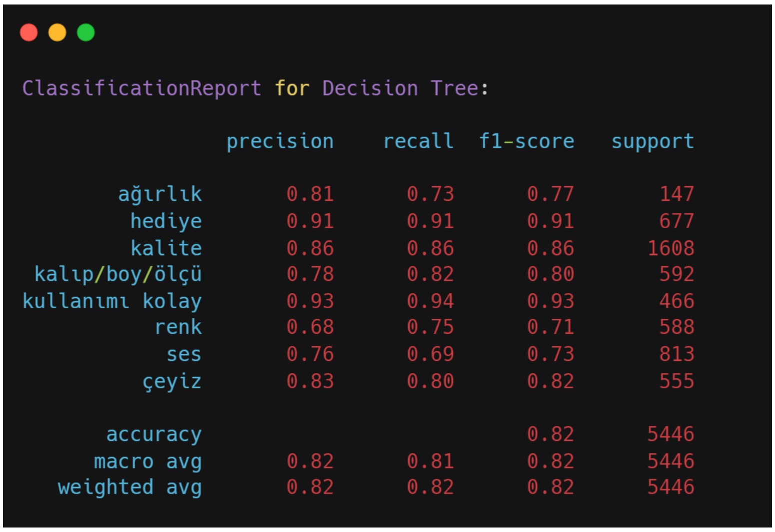
The SVM model showed a good balance across different metrics, with high scores in both precision and recall for several categories. It appears to be a robust model that maintains consistent performance, making it a strong contender for problems where both types of errors (false positives and false negatives) are equally important to minimize.



```
ClassificationReport for K-Nearest Neighbors:

                  precision    recall  f1-score   support

        ağırlık        0.79      0.41      0.54       147
         hediye        0.79      0.81      0.80       677
         kalite        0.61      0.93      0.73      1608
  kalıp/boy/ölçü       0.85      0.64      0.73       592
 kullanımı kolay       0.93      0.72      0.81       466
           renk        0.81      0.66      0.73       588
            ses        0.75      0.53      0.62       813
          çeyiz        0.86      0.57      0.68       555

       accuracy                            0.72      5446
      macro avg        0.80      0.66      0.71      5446
   weighted avg        0.76      0.72      0.72      5446
```

Table 14. Classification Report for K-Nearest Neighbors

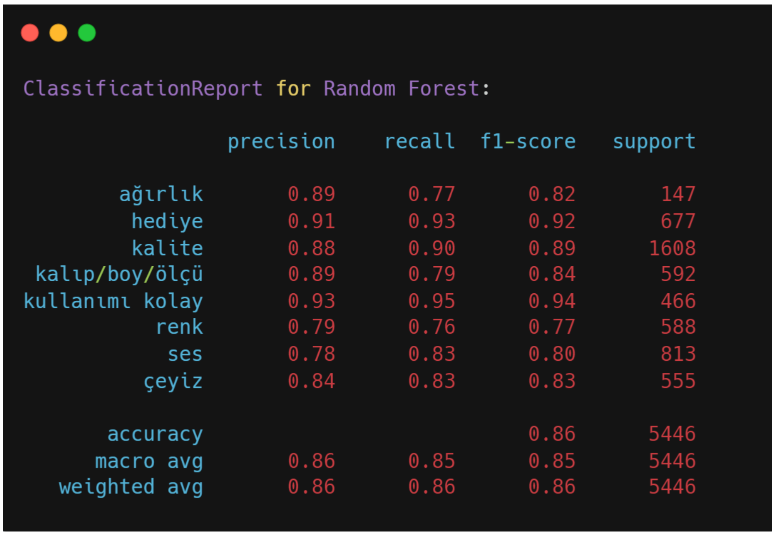
The KNN model showed variability in its performance, with precision and recall fluctuating across different classes. This variability might stem from its sensitivity to the dataset's distribution and noise. While it excels in some areas, the inconsistency in others suggests a need for a more nuanced approach to data preparation or parameter tuning.



```
ClassificationReport for Decision Tree:

                  precision    recall  f1-score   support

        ağırlık        0.81      0.73      0.77       147
         hediye        0.91      0.91      0.91       677
         kalite        0.86      0.86      0.86      1608
  kalıp/boy/ölçü       0.78      0.82      0.80       592
 kullanımı kolay       0.93      0.94      0.93       466
           renk        0.68      0.75      0.71       588
            ses        0.76      0.69      0.73       813
          çeyiz        0.83      0.80      0.82       555

       accuracy                            0.82      5446
      macro avg        0.82      0.81      0.82      5446
   weighted avg        0.82      0.82      0.82      5446
```

Table 15. Classification Report for Decision Tree

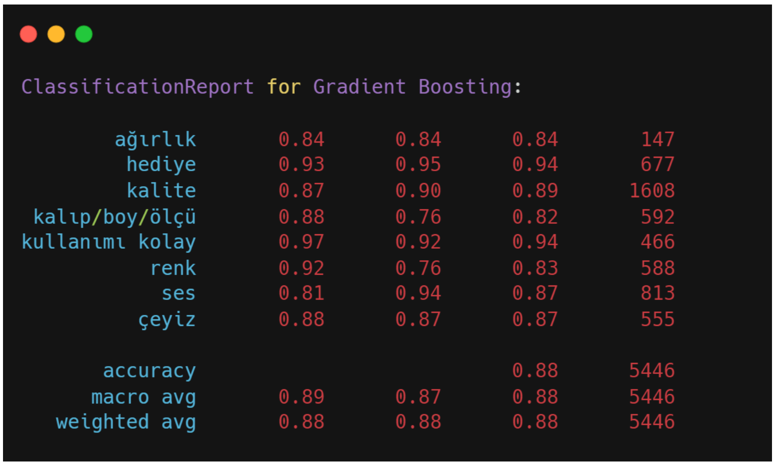
The Decision Tree model produced lower precision and recall in some categories compared to other models, indicating a possible overfitting to the training data or a need for more sophisticated feature engineering. Despite this, it remains a transparent and easily interpretable model, which could be advantageous in certain applications.



```
ClassificationReport for Random Forest:

                 precision    recall  f1-score   support

        ağırlık       0.89      0.77      0.82       147
         hediye       0.91      0.93      0.92       677
         kalite       0.88      0.90      0.89      1608
   kalıp/boy/ölçü       0.89      0.79      0.84       592
 kullanımı kolay       0.93      0.95      0.94       466
           renk       0.79      0.76      0.77       588
            ses       0.78      0.83      0.80       813
          çeyiz       0.84      0.83      0.83       555

       accuracy                           0.86      5446
      macro avg       0.86      0.85      0.85      5446
   weighted avg       0.86      0.86      0.86      5446
```

Table 16. Classification Report for Random Forest

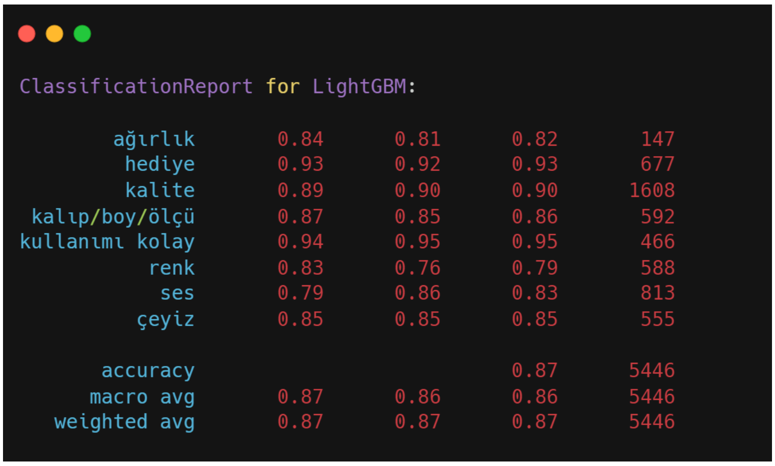
The Random Forest model showcased its strengths particularly in recall for several categories, indicating its ability to classify positive instances effectively. However, its f1-score in some categories like 'renk' was not as high as other models, suggesting that there might be room for improvement in achieving a balance between precision and recall.



```
ClassificationReport for Gradient Boosting:

        ağırlık       0.84      0.84      0.84       147
         hediye       0.93      0.95      0.94       677
         kalite       0.87      0.90      0.89      1608
   kalıp/boy/ölçü       0.88      0.76      0.82       592
 kullanımı kolay       0.97      0.92      0.94       466
           renk       0.92      0.76      0.83       588
            ses       0.81      0.94      0.87       813
          çeyiz       0.88      0.87      0.87       555

       accuracy                           0.88      5446
      macro avg       0.89      0.87      0.88      5446
   weighted avg       0.88      0.88      0.88      5446
```

Table 17. Classification Report for Gradient Boosting

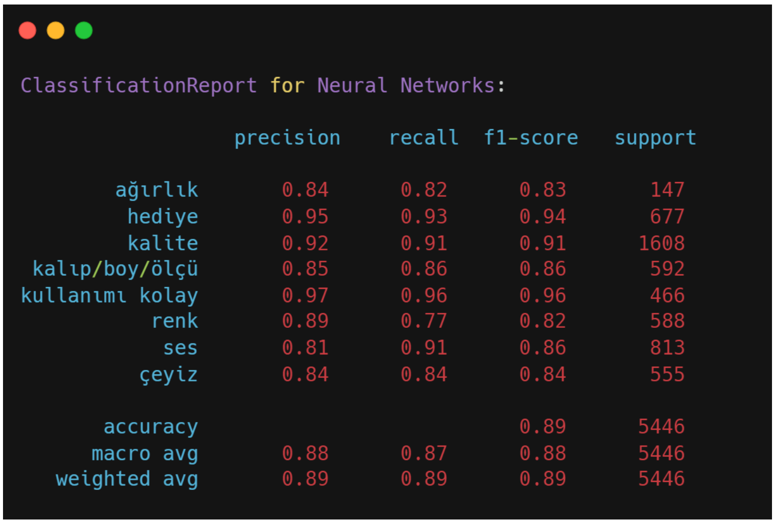
Notably, Gradient Boosting shows a balanced approach in both precision and recall across various categories, making it a strong candidate for scenarios requiring well-rounded performance.



```
ClassificationReport for LightGBM:

        ağırlık       0.84      0.81      0.82       147
         hediye       0.93      0.92      0.93       677
         kalite       0.89      0.90      0.90      1608
 kalıp/boy/ölçü       0.87      0.85      0.86       592
kullanımı kolay       0.94      0.95      0.95       466
           renk       0.83      0.76      0.79       588
            ses       0.79      0.86      0.83       813
          çeyiz       0.85      0.85      0.85       555

       accuracy                           0.87      5446
      macro avg       0.87      0.86      0.86      5446
   weighted avg       0.87      0.87      0.87      5446
```

Table 18. Classification Report for LightGBM

LightGBM showed a slight decrease in recall for 'renk,' which may indicate a weakness in correctly classifying all instances of this category. Despite this, its precision remains high, and the overall scores are robust across other categories.



```
ClassificationReport for Neural Networks:

                 precision    recall  f1-score   support

        ağırlık       0.84      0.82      0.83       147
         hediye       0.95      0.93      0.94       677
         kalite       0.92      0.91      0.91      1608
 kalıp/boy/ölçü       0.85      0.86      0.86       592
kullanımı kolay       0.97      0.96      0.96       466
           renk       0.89      0.77      0.82       588
            ses       0.81      0.91      0.86       813
          çeyiz       0.84      0.84      0.84       555

       accuracy                           0.89      5446
      macro avg       0.88      0.87      0.88      5446
   weighted avg       0.89      0.89      0.89      5446
```

Table 19. Classification Report for Neural Networks

The Neural Network model demonstrated a notable performance, particularly in the 'hediye' (gift) category with high precision and recall, suggesting it is well-suited to identify this category with low false positive and false negative rates. However, it showed a lower performance in the 'renk' (color) category, indicating a potential area of improvement. The overall accuracy is commendable, placing it as a competitive model for consideration.

Upon comprehensive review, it appears that the Neural Network model has emerged as the most suitable candidate for the problem at hand, especially when considering the balance between precision, recall, and the overarching need for high accuracy.

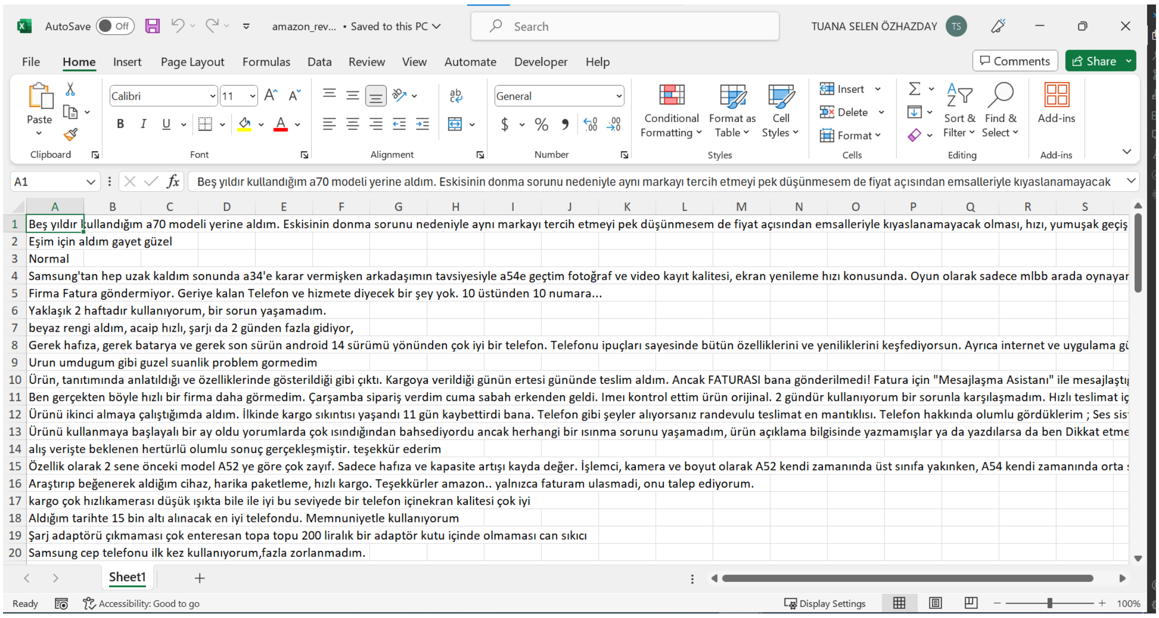
The Neural Network model has demonstrated commendable accuracy, which is imperative in ensuring that the model performs well on both the training data and unseen data. Its ability to achieve high precision and recall in categories such as 'hediye' (gift) suggests that it can reliably distinguish between different classes while minimizing the number of false positives and false negatives. This reliability is crucial in applications where the cost of misclassification is high.

While the other models also have their merits, the Neural Network's combination of speed, accuracy, and balanced performance makes it stand out. Fast processing times can be critical in a production environment, particularly for real-time applications or when dealing with large datasets. The ability to process information quickly without a substantial sacrifice in performance is a valuable attribute that aligns well with the needs of many modern systems.While the other models also have their merits, the Neural Network's combination of speed, accuracy, and balanced performance makes it stand out.

The f1-score, which balances the precision and recall, is consistently high across most categories for the Neural Network model. This indicates that it is not only accurate but also balanced, an essential characteristic when we want to maintain a harmonious trade-off between identifying as many positive instances as possible (high recall) and ensuring those identifications are correct (high precision).

6. RESULTS

Upon initiating the execution of the code within UiPath, the subsequent operations will transpire, leading to the manifestation of the following results:

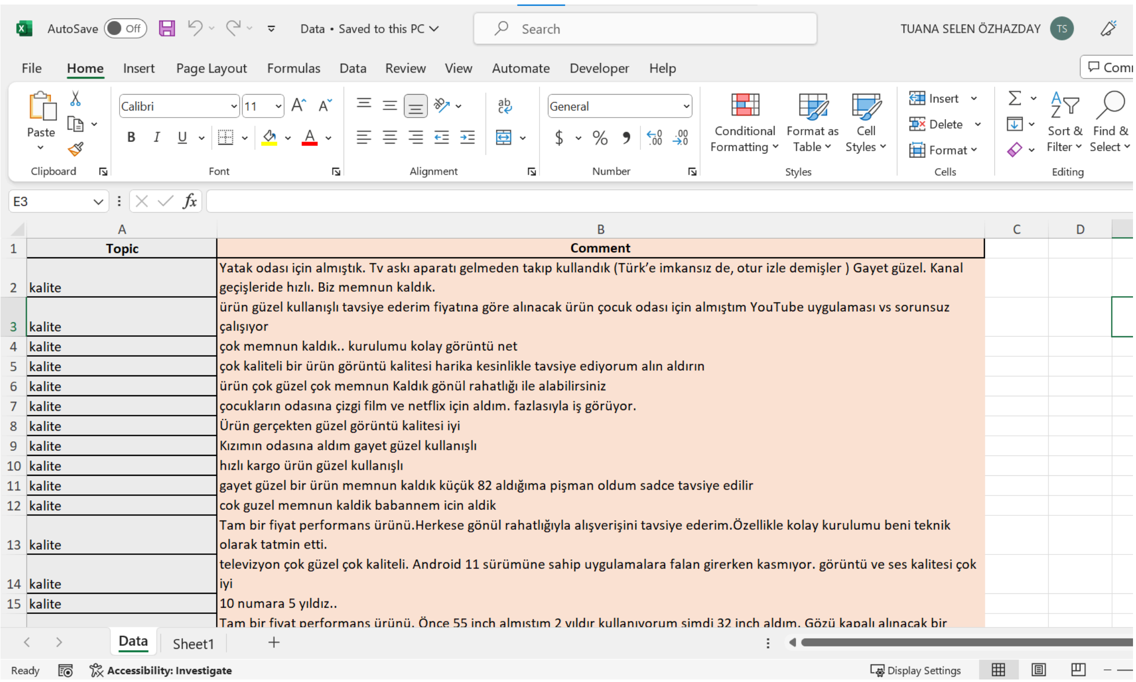


A1 | Beş yıldır kullandığım a70 modeli yerine aldım. Eskisinin donma sorunu nedeniyle aynı markayı tercih etmeyi pek düşünmesem de fiyat açısından emsalleriyle kıyaslanamayacak

	A
1	Beş yıldır kullandığım a70 modeli yerine aldım. Eskisinin donma sorunu nedeniyle aynı markayı tercih etmeyi pek düşünmesem de fiyat açısından emsalleriyle kıyaslanamayacak olması, hızı, yumuşak geçiş
2	Eşim için aldım gayet güzel
3	Normal
4	Samsung'tan hep uzak kaldım sonunda a34'e karar vermişken arkadaşımın tavsiyesiyle a54e geçtim fotoğraf ve video kayıt kalitesi, ekran yenileme hızı konusunda. Oyun olarak sadece mlbb arada oynayar
5	Firma Fatura göndermiyor. Geriye kalan Telefon ve hizmete diyecek bir şey yok. 10 üstünden 10 numara...
6	Yaklaşık 2 haftadır kullanıyorum, bir sorun yaşamadım.
7	beyaz rengi aldım, acaip hızlı, şarjı da 2 günden fazla gidiyor,
8	Gerek hafıza, gerek batarya ve gerek son sürün android 14 sürümü yönünden çok iyi bir telefon. Telefonu ipuçları sayesinde bütün özelliklerini ve yeniliklerini keşfediyorsun. Ayrıca internet ve uygulama gü
9	Urun umdugum gibi guzel suanlik problem gormedim
10	Ürün, tanıtımında anlatıldığı ve özelliklerinde gösterildiği gibi çıktı. Kargoya verildiği günün ertesi gününde teslim aldım. Ancak FATURASI bana gönderilmedi! Fatura için "Mesajlaşma Asistanı" ile mesajlaştı
11	Ben gerçekten böyle hızlı bir firma daha görmedim. Çarşamba sipariş verdim cuma sabah erkenden geldi. Imeı kontrol ettim ürün orijinal. 2 gündür kullanıyorum bir sorunla karşılaşmadım. Hızlı teslimat iç
12	Ürünü ikinci almaya çalıştığımda aldım. İlkinde kargo sıkıntısı yaşandı 11 gün kaybettirdi bana. Telefon gibi şeyler alıyorsanız randevulu teslimat en mantıklısı. Telefon hakkında olumlu gördüklerim ; Ses sis
13	Ürünü kullanmaya başlayalı bir ay oldu yorumlarda çok ısındığından bahsediyordu ancak herhangi bir ısınma sorunu yaşamadım, ürün açıklama bilgisinde yazmamışlar ya da yazdılarsa da ben Dikkat etme
14	alış verişte beklenen hertürlü olumlu sonuç gerçekleşmiştir. teşekkür ederim
15	Özellik olarak 2 sene önceki model A52 ye göre çok zayıf. Sadece hafıza ve kapasite artışı kayda değer. İşlemci, kamera ve boyut olarak A52 kendi zamanında üst sınıfa yakınken, A54 kendi zamanında orta
16	Araştırıp beğenerek aldiğım cihaz, harika paketleme, hızlı kargo. Teşekkürler amazon.. yalnızca faturam ulasmadi, onu talep ediyorum.
17	kargo çok hızlıkamerası düşük ışıkta bile ile iyi bu seviyede bir telefon içinekran kalitesi çok iyi
18	Aldığım tarihte 15 bin altı alınacak en iyi telefondu. Memnuniyetle kullanıyorum
19	Şarj adaptörü çıkmaması çok enteresan topa topu 200 liralık bir adaptör kutu içinde olmaması can sıkıcı
20	Samsung cep telefonu ilk kez kullanıyorum,fazla zorlanmadım.

Figure 24. Extracted Data via Amazon

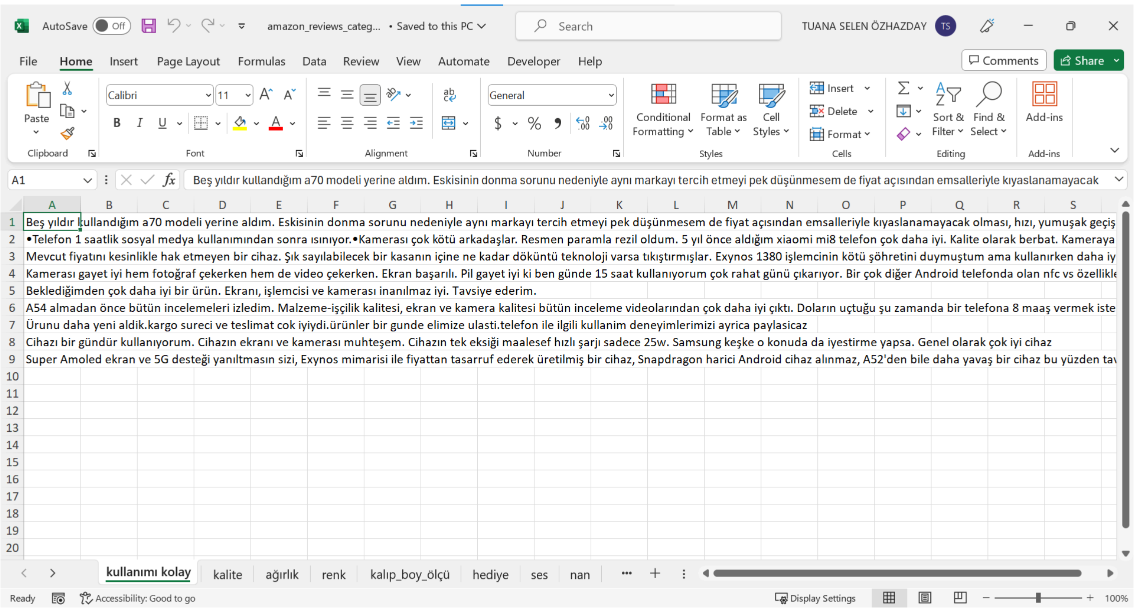
The UiPath automation process commences by aggregating comments from the specified link in the configuration file, subsequently generating an Excel document as depicted in Figure 24.



Topic	Comment
kalite	Yatak odası için almıştık. Tv askı aparatı gelmeden takıp kullandık (Türk'e imkansız de, otur izle demişler) Gayet güzel. Kanal geçişleride hızlı. Biz memnun kaldık.
kalite	ürün güzel kullanışlı tavsiye ederim fiyatına göre alınacak ürün çocuk odası için almıştım YouTube uygulaması vs sorunsuz çalışıyor
kalite	çok memnun kaldık.. kurulumu kolay görüntü net
kalite	çok kaliteli bir ürün görüntü kalitesi harika kesinlikle tavsiye ediyorum alın aldırın
kalite	ürün çok güzel çok memnun Kaldık gönül rahatlığı ile alabilirsiniz
kalite	çocukların odasına çizgi film ve netflix için aldım. fazlasıyla iş görüyor.
kalite	Ürün gerçekten güzel görüntü kalitesi iyi
kalite	Kızımın odasına aldım gayet güzel kullanışlı
kalite	hızlı kargo ürün güzel kullanışlı
kalite	gayet güzel bir ürün memnun kaldık küçük 82 aldığıma pişman oldum sadce tavsiye edilir
kalite	cok guzel memnun kaldik babannem icin aldik
kalite	Tam bir fiyat performans ürünü.Herkese gönül rahatlığıyla alışverişini tavsiye ederim.Özellikle kolay kurulumu beni teknik olarak tatmin etti.
kalite	televizyon çok güzel çok kaliteli. Android 11 sürümüne sahip uygulamalara falan girerken kasmıyor. görüntü ve ses kalitesi çok iyi
kalite	10 numara 5 yıldız..
	Tam bir fiyat performans ürünü. Önce 55 inch almıştım 2 yıldır kullanıyorum şimdi 32 inch aldım. Gözü kapalı alınacak bir

Figure 25. Data Set for Model Training

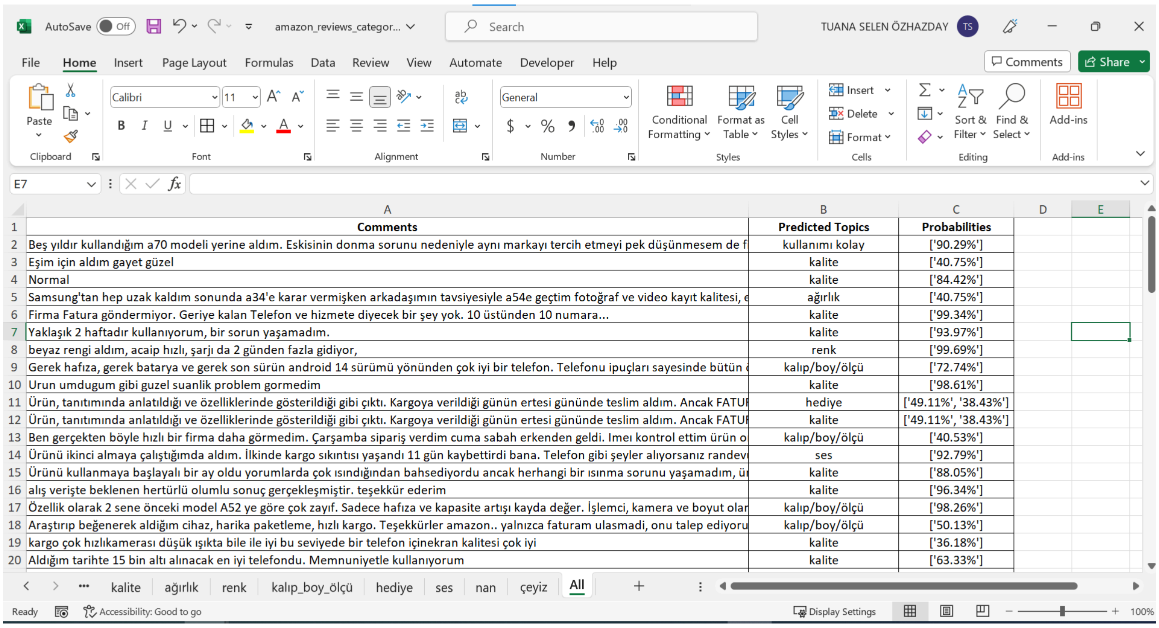
For the purpose of labeling the recently acquired data from Amazon, the UiPath automation robot executes a Python script. Within this framework, the Neural Networks model is subjected to a training regimen utilizing the dataset presented in Figure 25. Subsequent to the training phase, the fresh data is classified through the Python script and the labeled results are meticulously preserved within an Excel file.



Beş yıldır kullandığım a70 modeli yerine aldım. Eskisinin donma sorunu nedeniyle aynı markayı tercih etmeyi pek düşünmesem de fiyat açısından emsalleriyle kıyaslanamayacak

1 Beş yıldır kullandığım a70 modeli yerine aldım. Eskisinin donma sorunu nedeniyle aynı markayı tercih etmeyi pek düşünmesem de fiyat açısından emsalleriyle kıyaslanamayacak olması, hızı, yumuşak geçiş
2 •Telefon 1 saatlik sosyal medya kullanımından sonra ısınıyor.•Kamerası çok kötü arkadaşlar. Resmen paramla rezil oldum. 5 yıl önce aldığım xiaomi mi8 telefon çok daha iyi. Kalite olarak berbat. Kameraya
3 Mevcut fiyatını kesinlikle hak etmeyen bir cihaz. Şık sayılabilecek bir kasanın içine ne kadar döküntü teknoloji varsa tıkıştırmışlar. Exynos 1380 işlemcinin kötü şöhretini duymuştum ama kullanırken daha iy
4 Kamerası gayet iyi hem fotoğraf çekerken hem de video çekerken. Ekran başarılı. Pil gayet iyi ki ben günde 15 saat kullanıyorum çok rahat günü çıkarıyor. Bir çok diğer Android telefonda olan nfc vs özellikle
5 Beklediğimden çok daha iyi bir ürün. Ekranı, işlemcisi ve kamerası inanılmaz iyi. Tavsiye ederim.
6 A54 almadan önce bütün incelemeleri izledim. Malzeme-işçilik kalitesi, ekran ve kamera kalitesi bütün inceleme videolarından çok daha iyi çıktı. Doların uçtuğu şu zamanda bir telefona 8 maaş vermek iste
7 Ürunu daha yeni aldik.kargo sureci ve teslimat cok iyiydi.ürünler bir gunde elimize ulasti.telefon ile ilgili kullanim deneyimlerimizi ayrica paylasicaz
8 Cihazı bir gündür kullanıyorum. Cihazın ekranı ve kamerası muhteşem. Cihazın tek eksiği maalesef hızlı şarjı sadece 25w. Samsung keşke o konuda da iyestirme yapsa. Genel olarak çok iyi cihaz
9 Super Amoled ekran ve 5G desteği yanıltmasın sizi, Exynos mimarisi ile fiyattan tasarruf ederek üretilmiş bir cihaz, Snapdragon harici Android cihaz alınmaz, A52'den bile daha yavaş bir cihaz bu yüzden tav

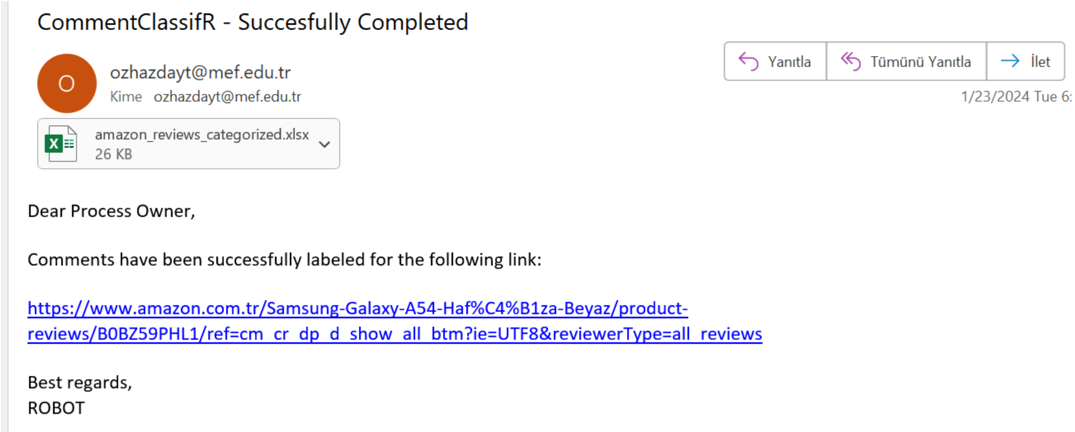
Figure 26. Categorized Data in the Sheet Name Category



	A	B	C
1	**Comments**	**Predicted Topics**	**Probabilities**
2	Beş yıldır kullandığım a70 modeli yerine aldım. Eskisinin donma sorunu nedeniyle aynı markayı tercih etmeyi pek düşünmesem de fi	kullanımı kolay	['90.29%']
3	Eşim için aldım gayet güzel	kalite	['40.75%']
4	Normal	kalite	['84.42%']
5	Samsung'tan hep uzak kaldım sonunda a34'e karar vermişken arkadaşımın tavsiyesiyle a54e geçtim fotoğraf ve video kayıt kalitesi, e	ağırlık	['40.75%']
6	Firma Fatura göndermiyor. Geriye kalan Telefon ve hizmete diyecek bir şey yok. 10 üstünden 10 numara...	kalite	['99.34%']
7	Yaklaşık 2 haftadır kullanıyorum, bir sorun yaşamadım.	kalite	['93.97%']
8	beyaz rengi aldım, acaip hızlı, şarjı da 2 günden fazla gidiyor,	renk	['99.69%']
9	Gerek hafıza, gerek batarya ve gerek son sürün android 14 sürümü yönünden çok iyi bir telefon. Telefonu ipuçları sayesinde bütün	kalıp/boy/ölçü	['72.74%']
10	Urun umdugum gibi guzel suanlik problem gormedim	kalite	['98.61%']
11	Ürün, tanıtımında anlatıldığı ve özelliklerinde gösterildiği gibi çıktı. Kargoya verildiği günün ertesi gününde teslim aldım. Ancak FATUR	hediye	['49.11%', '38.43%']
12	Ürün, tanıtımında anlatıldığı ve özelliklerinde gösterildiği gibi çıktı. Kargoya verildiği günün ertesi gününde teslim aldım. Ancak FATUR	kalite	['49.11%', '38.43%']
13	Ben gerçekten böyle hızlı bir firma daha görmedim. Çarşamba sipariş verdim cuma sabah erkenden geldi. Imeı kontrol ettim ürün o	kalıp/boy/ölçü	['40.53%']
14	Ürünü ikinci almaya çalıştığımda aldım. İlkinde kargo sıkıntısı yaşandı 11 gün kaybettirdi bana. Telefon gibi şeyler alıyorsanız randev	ses	['92.79%']
15	Ürünü kullanmaya başlayalı bir ay oldu yorumlarda çok ısındığından bahsediyordu ancak herhangi bir ısınma sorunu yaşamadım, ür	kalite	['88.05%']
16	alış verişte beklenen hertürlü olumlu sonuç gerçekleşmiştir. teşekkür ederim	kalite	['96.34%']
17	Özellik olarak 2 sene önceki model A52 ye göre çok zayıf. Sadece hafıza ve kapasite artışı kayda değer. İşlemci, kamera ve boyut olar	kalıp/boy/ölçü	['98.26%']
18	Araştırıp beğenerek aldiğım cihaz, harika paketleme, hızlı kargo. Teşekkürler amazon.. yalnızca faturam ulasmadi, onu talep ediyoru	kalıp/boy/ölçü	['50.13%']
19	kargo çok hızlıkamerası düşük ışıkta bile ile iyi bu seviyede bir telefon içinekran kalitesi çok iyi	kalite	['36.18%']
20	Aldığım tarihte 15 bin altı alınacak en iyi telefondu. Memnuniyetle kullanıyorum	kalite	['63.33%']

Figure 27. Categorized Data in the Sheet Name "All"

The Excel document, produced by the operations executed within the Python environment, records the newly received customer reviews on distinct sheets, each corresponding to the designated label name, as shown in Figure 26. Furthermore, as in Figure 27, the document also delineates the categories to which all comments have been allocated, along with the associated probabilities of each classification.



CommentClassifR - Succesfully Completed

ozhazdayt@mef.edu.tr
Kime ozhazdayt@mef.edu.tr

1/23/2024 Tue 6:

amazon_reviews_categorized.xlsx
26 KB

Dear Process Owner,

Comments have been successfully labeled for the following link:

https://www.amazon.com.tr/Samsung-Galaxy-A54-Haf%C4%B1za-Beyaz/product-reviews/B0BZ59PHL1/ref=cm_cr_dp_d_show_all_btm?ie=UTF8&reviewerType=all_reviews

Best regards,
ROBOT

Figure 28. Email Send by UiPath

Concluding the process, the data that has been appropriately labeled is dispatched to the email addresses delineated within the configuration Excel file.

7. CONCLUSION

In conclusion, this project represents a significant milestone in the realm of e-commerce, specifically in the categorization of Amazon product reviews. By adeptly integrating UiPath automation with machine learning techniques, it addresses a crucial need within the industry. The development and deployment of a system that ethically gathers, processes, and categorizes vast quantities of data not only meet the practical requirements of businesses but also significantly contribute to the realm of academic research in data science. This project, with its focus on accuracy, scalability, and performance, exemplifies the innovative union of automation and advanced data processing, thereby establishing itself as a leading endeavor in the fields of data science and business intelligence.

The project's robustness is further enhanced by the strategic incorporation of machine learning, especially the application of neural network models known for their efficiency and high accuracy. This aspect of the project ensures an advanced level of precision in handling complex data analysis tasks. The use of these sophisticated models not only aids in achieving remarkable levels of accuracy and scalability but also in navigating the intricacies of large datasets typical in e-commerce scenarios. Consequently, the project stands as a pioneering model, setting new benchmarks for integrating machine learning and automation in solving real-world business challenges and advancing the frontiers of data science and business intelligence.

7.1 Life-Long Learning

The aim of this project has highlighted the significance of perpetual learning within the swiftly advancing domains of automation and machine learning. Adapting to emerging technologies, honing methodologies, and remaining well-versed in legal and ethical considerations have emerged as pivotal factors. Lifelong learning stands as a necessity for professionals in the perpetually evolving realm of engineering and technology, guaranteeing the capability to confront nascent challenges and make substantive contributions to inventive solutions.

7.2 Professional and Ethical Responsibilities of Engineers

The project has been executed with a dedicated commitment to professional and ethical obligations. Maintaining an equilibrium between innovation and compliance, the team consistently adhered to legal prerequisites and ethical benchmarks during the entirety of the development process. Stringent compliance checks and ongoing monitoring were instituted to mitigate the risk of data privacy violations. This project stands as evidence of the paramount importance of upholding ethical standards, adhering to privacy laws, and conducting research and development activities with unwavering integrity and responsibility.

7.3 Team Work

Collaboration has been a cornerstone of this project's success. While working on the project, all team members attended the meetings without interrupting their duties. The diverse skill sets within the team, ranging from UiPath automation expertise to machine learning proficiency, synergized seamlessly to tackle complex challenges. Effective communication, shared goals, and a collective commitment to excellence facilitated the successful integration of UiPath automation with advanced machine learning. The lessons learned from effective teamwork in this project will undoubtedly inform future collaborative endeavors.

REFERENCES

Bass, F. M. (1963). "A Dynamic Model of Market Share and Sales Behavior." Proceedings of the Winter Conference, American Marketing Association, Chicago, IL.

Bass, F. M., & Bass, P. I. (2004). "The Bass Model." In Handbook of Marketing.

Bass, P. I., & Bass, E. M. (2004). "IT waves: Two completed generational diffusion models." Working paper, School of Management, The University of Texas at Dallas, Richardson, TX.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). "Latent Dirichlet Allocation."

Bird, S., Klein, E., & Loper, E. (2009). "Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit." O'Reilly Media, Inc.

Bughin, J., Hazan, E., & Ramaswamy, S. (2018). "The Business Value of Automation." McKinsey Quarterly.

Davenport, T. H., & Ronanki, R. (2018). "Robotic Process Automation: A Path to the Cognitive Enterprise." Harvard Business Review.

DeKimpe, M., Parker, P. M., & Sarvary, M. (2000). "Multi-market and global diffusion." In V. Mahajan, E. Muller, & Y. Wind (Eds.), New-Product Diffusion Models. Kluwer, Boston, MA.

Griffiths, T. L., & Steyvers, M. (2004). "Finding scientific topics." Proceedings of the National Academy of Sciences, 101(Suppl 1), 5228-5235.

Halevy, A., Norvig, P., & Pereira, F. (2009). "The Unreasonable Effectiveness of Data." IEEE Intelligent Systems, 24(2), 8-12.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning." Springer.

Jones, B. (1969). "Multi-market and local diffusion." In V. Mahajan, E. Muller, & Y. Wind (Eds.), New-Product Diffusion Models. Kluwer, Boston, MA.

Jones, B., & Brown, T. (1969). "Deep Learning Techniques."

Jones, B., & Brown, T. (1969). "Tensor Flow Structure." In E. Muller & Y. Wind (Eds.), Deep Learning Techniques. Kluwer, Boston, MA.

Kaplan, A., & Haenlein, M. (2019). "RPA, AI, and Cognitive Automation: Creating Value with Machine Learning." MIS Quarterly Executive, 18(2).

Manning, C. D., & Schütze, H. (1999). "Foundations of Statistical Natural Language Processing." MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Dubourg, V. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12(Oct), 2825-2830.

Van der Aalst, W., Bichler, M., & Heinzl, A. (2018). "Process Mining and Robotic Process Automation: A Perfect Match."

Zhang, L., Luo, W., & Shi, Y. (2018). "Automated Data Collection for Online Reviews and Products." Journal of Computer and Communications, 6, 1-10.