CUSTOMER SATISFACTION ANALYSIS OF PRODUCT REVIEWS IN ONLINE SHOPPING PLATFORMS USING SUPPORT VECTOR MACHINE

This CS Thesis Writing Presented to the
Faculty of the College of Computer Studies
AMA Computer College

In Partial Fulfillment of the Requirements of the Degree of
Bachelor of Science in Computer Science (BSCS)

By

Bawalan, Krishna Samantha A.
Hernandez, John Jester C.
Quitan, Elaine M.

December 2023

APPROVAL SHEET

This CS Thesis Writing project entitled ***"CUSTOMER SATISFACTION ANALYSIS OF PRODUCT REVIEWS IN ONLINE SHOPPING PLATFORMS USING SUPPORT VECTOR MACHINE*"**, prepared and submitted by KRISHNA SAMANTHA A. BAWALAN, JOHN JESTER S. HERNANDEZ, ELAINE M. QUITAN in partial fulfillment of the course requirements for the degree of BACHELOR OF SCIENCE IN COMPUTER SCIENCE, has been examined and recommended for acceptance and approval for Oral Examination.

Jacinto
Angela Elaine L. Jacinto
ANGELA ELAINE LUCIDO JACINTO
Adviser

Approved by the Committee of Oral Examination on December 13, 2023
with the grade of ____.

____
Chairman

____
Member Panel 1

____
Member Panel 2

Accepted as partial fulfillment of the course requirements of the degree in BACHELOR OF SCIENCE IN COMPUTER SCIENCE

DIARECSON RIVERA
School Director – AMACC Cavite
Date:____________

ACKNOWLEDGEMENTS

We wish to convey our heartfelt appreciation to our undergraduate thesis adviser, Ma'am Angela Elaine L. Jacinto. Her willingness to take us under her guidance and assist us in crafting and structuring our manuscript is something we deeply cherish.

Additionally, we would like to express our gratitude to Dr. Regene G. Hernandez. Despite our abrupt approach, she provided unwavering support, valuable guidance, and remarkable patience throughout our thesis journey, especially when the time when we don’t know what to do and how to go on with the thesis. Her patience in answering our questions and concerns in her busy times is so much appreciated. And her expertise and encouragement played a pivotal role in enhancing the quality of our project and the content of our manuscript.

Our thanks also extend to Sir Mosiah David A. Yacapin, who aided us in formulating our thesis title. His generous assistance, despite his busy schedule and absence of official responsibility, was a driving force that kept us committed to this thesis.

We also wish to express our gratitude to the panelists who participated in each of our defense sessions. Their critiques, comments, commendations, and suggestions regarding our manuscript and project have been invaluable. Their insights not only guided us in refining our work but also encouraged us to explore alternative avenues for improvement.

We would like to convey our appreciation to Sir Diarecson A. Rivera, the School Director at AMACC - Cavite. Sir Rivera consistently supported our requests for access to campus facilities, including rooms for our thesis project development and defense preparations. We are thankful for his prompt and positive responses, even during the busiest periods. His assistance was instrumental in facilitating our work and ensuring its smooth progression.

We would like to express our gratitude to each member of our team. Each one of us dedicated our best efforts to fulfill our respective tasks and made a conscious effort to appreciate and understand the unique circumstances of our team members. This mutual respect for one another's contributions was instrumental in the success of this project. Together, we've truly made a significant impact.

Furthermore, we express our appreciation to our classmates, colleagues, friends, and acquaintances who generously shared their expertise and experiences, answering our questions and providing assistance in their respective fields. Their patience and willingness to offer advice amidst their own demanding responsibilities are deeply touching.

Lastly, we're grateful to our families for their unwavering encouragement and belief in our abilities. Their emotional support, love and understanding provided us with the strength to persevere throughout this thesis project.

This thesis would not have been possible without the contributions and support of these remarkable individuals and organizations. We extend our heartfelt thanks to each and every one of them.

ABSTRACT

This thesis study delves into creating and implementing an effective customer satisfaction analysis system tailored for online shopping platforms. Through careful exploration and decision-making, the study introduces a Support Vector Machine model skilled at categorizing customer satisfaction into positive, negative, and neutral. Using a thoughtful approach to dataset sampling, the study optimizes training efficiency while considering resource constraints. The evaluation process reveals that training the model on 30% of the dataset strikes a commendable balance between accuracy and computational resources. The model's effectiveness is further demonstrated through its application to diverse datasets from platforms like eBay, Flipkart, Sephora, and Shopee, showcasing commendable accuracy rates. The study culminates in the development of a user-centric web application, offering individual and consolidated review analyses, coupled with insightful keyword extraction. The web application stands as a clear demonstration of the study's commitment to advancing decision support for small businesses by extracting invaluable insights from customer reviews. The thesis concludes with recommendations for scaling up review processing capacity, exploring advanced machine learning models, diversifying training datasets, and fostering continuous improvement through user feedback, collectively aimed at enhancing the system's effectiveness and user experience.

TABLE OF CONTENTS

APPROVAL SHEET .... **ii**
ACKNOWLEDGEMENTS .... **iii**
ABSTRACT .... **v**
TABLE OF CONTENTS .... **vi**
LIST OF TABLES .... **viii**
LIST OF FIGURES .... **ix**
LIST OF APPENDICES .... **x**
LIST OF ABBREVIATIONS AND SYMBOLS .... **xi**
DEFINITION OF TERMS .... **xii**
CHAPTER I **1**
- **Project Context** 1
- **Purpose and Description** 3
- **Objectives of the Study** 4
- **Significance of the Study** 4
- **Scope and Limitation** 5

CHAPTER II **8**
- **Related Literature** 8
  - **Foreign Literature** 8
  - **Local Literature** 10
- **Related Studies** 12
  - **Foreign Studies** 12
  - **Local Studies** 17
- **Synthesis** 19
- **Technical Background** 23

CHAPTER III **25**
- **Software Design** 25
  - **Requirements Gathering** 27
  - **System Design** 27
  - **Implementation** 28
  - **Testing and Validation** 28
  - **Deployment** 28

Maintenance 29
Conceptual Design 29
Application Requirement 33
System Architecture/System Flow 35
Input and Output Reports and Analysis 41
Multiple Constraints 43
Algorithm Use 44
CHAPTER IV 48
A. Model Training and Evaluation 48
B. Model Verification 53
C. Developed Web Application 55
CHAPTER V 62
Conclusion 62
Recommendations 64
REFERENCES 67
APPENDICES 74

LIST OF TABLES

Table 1. Equivalent File Sizes per Dataset Usage Percentage 49

Table 2. Model Evaluation per Dataset Percentage 50

LIST OF FIGURES

Figure 1. Modified Waterfall Model 26

Figure 2. The Conceptual Framework 29

Figure 3. Simplified Use Case Diagram 30

Figure 4. Detailed Use Case Diagram 31

Figure 5. System flow of machine learning model training 35

Figure 6. TF-IDF Formula 38

Figure 7. Dataset Input to Model Evaluation Output 41

Figure 8. User Input to Classifier Output 42

Figure 9. Visualization of Project Constraints 43

Figure 10. LinearSVC for Multi Classification 46

Figure 11. LinearSVC Hyperplane in Multi Classification using OVR Approach . 47

Figure 12. Impact of Dataset Size on Accuracy and Resource Utilization 51

Figure 13. Home Page 56

Figure 14. Features Page 57

Figure 15. Context Page 58

Figure 16. Meet the Team Page 59

Figure 17. Individual Product Review Classifier Page 60

Figure 18. Consolidated Product Review Analyzer Page 60

Figure 19. Footer Part for Disclaimer 61

LIST OF APPENDICES

Appendix A. 1-Dataset_Prep.ipynb 74

Appendix B. data_preprocess.py 76

Appendix C. 2-SVM_ModelTrain.ipynb 76

LIST OF ABBREVIATIONS AND SYMBOLS

AI	Artificial Intelligence
CSS	Cascading Style Sheets
HTML	HyperText Markup Language
IDE	Integrated Development Environment
LR	Logistic Regression
LTSM	Long Short-Term Memory
ML	Machine Learning
MNB	Multinomial Naive Bayes
NLP	Natural Language Processing
OVR	one-vs-rest
SDLC	Software Development Life Cycle
SVC	Support Vector Classifier
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
VADER	Valence Aware Dictionary for Sentiment Reasoning

DEFINITION OF TERMS

Accuracy. A measure of the correctness of a machine learning model's predictions indicates the proportion of correctly classified instances out of the total.

Classifier. A machine learning algorithm or model that is trained to categorize data points into predefined classes or categories based on their features.

Customer feedback. Customer feedback refers to the opinions, comments, and sentiments expressed by users of a product or service.

Customer Satisfaction levels. Customer satisfaction levels represent the degree of contentment or fulfillment experienced by customers with a particular product or service.

Data analysis. The process of inspecting, cleaning, transforming, and interpreting data to discover useful information and make informed decisions.

Data preprocessing. The initial stage of data preparation, involving tasks like cleaning, formatting, and transforming raw data into a usable format for analysis.

Data quality. The measure of the reliability, accuracy, and consistency of data, which impacts the validity of insights drawn from it.

Data Science. An interdisciplinary field that uses scientific methods, algorithms, processes, and systems to extract knowledge and insights from structured and unstructured data.

Data visualization. The graphical representation of data to aid in understanding patterns, trends, and relationships within the data.

Dataset. A structured collection of data used for analysis, typically organized into rows (samples) and columns (features).

Dataset splitting. The process of dividing a dataset into two or more subsets, often a training set and a testing set, to evaluate machine learning models.

Decision boundary. In machine learning, the boundary that separates different classes or categories in a classification problem.

Decision Support System. A computer-based system that aids decision-makers in solving complex problems by providing relevant information and analysis.

Deployment server. A server where machine learning models or applications are hosted and made available for use by end-users.

E-Commerce. Electronic commerce is the buying and selling of goods or services over the Internet.

Evaluation matrix. A structured way to assess and measure the performance of a system, model, or process, often including metrics such as accuracy, precision, recall, and F-measure.

F-measure. A metric that combines precision and recall to provide a single score for evaluating the performance of a classification model.

Feature extraction. The process of selecting and transforming relevant information (features) from data for use in machine learning models.

Feedback. Information or responses provided by users or customers, often used to make improvements or adjustments.

Hyperplane. In the context of Support Vector Machines (SVM), a hyperplane is a multidimensional surface that separates data into different classes.

Kernel. A function used in machine learning algorithms, such as SVM, to measure the similarity between data points.

Linear kernel. A kernel function that computes the dot product between data points, often used in linear SVM.

Linear SVC. Linear Support Vector Classification, a specific implementation of SVM for linearly separable data.

Machine learning. A subset of artificial intelligence (AI) that focuses on developing algorithms and models that enable computers to learn from and make predictions or decisions based on data.

Machine learning algorithm. A set of rules and procedures used by a machine learning model to make predictions or decisions.

Machine learning model. The result of training a machine learning algorithm on data that can be used to make predictions or classify new data points.

Margin. In SVM, the separation distance between the decision boundary (hyperplane) and the nearest data point of any class.

Model evaluation. The process of assessing the performance and effectiveness of a machine learning model using various metrics and techniques.

Model training. Model training involves the process of instructing a machine learning algorithm using a labeled dataset to enable it to make accurate predictions or classifications.

Model tuning. The iterative process of adjusting the parameters and settings of a machine learning model to optimize its performance.

Modified waterfall model. The modified waterfall model is a structured software development methodology that emphasizes sequential progression through defined phases.

Module. A self-contained software or hardware unit with a specific function or purpose within a larger system.

Multi-classification. A type of machine learning task where data is classified into more than two distinct categories or classes.

Natural Language Processing (NLP). The field of AI focused on the interaction between computers and human language, including tasks like text analysis and language understanding.

Negation. In natural language processing, the presence of words or phrases that indicate negation or the opposite meaning.

Online Shopping Platform. A digital platform or website that allows users to browse, select, and purchase products or services over the Internet.

Precision. A metric that measures the accuracy of positive predictions made by a classification model.

Processed data. Data that has been cleaned, transformed, and prepared for analysis or modeling.

Product reviews. Feedback or assessments provided by customers or users about products or services they have purchased or used.

Python. A popular programming language commonly used in data science, machine learning, and web development.

Rating score. The rating score represents a numerical evaluation assigned to a product or service by customers, often indicating the level of satisfaction.

Raw data. Raw data refers to unprocessed and unorganized information that is collected directly from sources.

Recall. Recall, in the context of sentiment analysis, refers to the ability of a model to correctly identify and retrieve instances of a particular sentiment from the dataset. It is a crucial metric for evaluating the model's performance in capturing relevant information.

Scikit-learn. An open-source machine learning library for Python that provides tools for data analysis and modeling.

Sentiment analysis. Sentiment analysis is a natural language processing (NLP) technique that involves determining and extracting sentiments or opinions expressed in text.

Sentiment classification. The task of categorizing text or data into sentiment categories, such as positive, neutral, or negative.

Software Development Life Cycle (SDLC). A structured approach to software development that includes planning, design, development, testing, deployment, and maintenance phases.

Stop words. Commonly used words (e.g., "the," "is," "and") that are often removed from text data during preprocessing to reduce noise.

Structured data. Data that is organized into a specific format, making it easy to query and analyze.

Support Vector Classifier (SVC). A machine learning algorithm, part of the Support Vector Machines (SVM) family, used for sentiment analysis. It establishes a decision boundary to categorize product reviews based on sentiments like positive, neutral, or negative.

Support Vector Machine (SVM). A machine learning algorithm used for classification and regression tasks, often employed for its effectiveness in high-dimensional spaces.

Term Frequency-Inverse Document Frequency (TF-IDF). A numerical statistic used to evaluate the importance of a word within a document relative to a collection of documents, often used in text analysis.

Testing data. Data used to evaluate the performance of a machine learning model, typically distinct from the training data.

Testing environment. The controlled setting where software or models are tested to ensure they meet specified requirements and function correctly.

Training data. Data used to train a machine learning model, typically consisting of labeled examples used for learning patterns.

Unstructured data. Data that lacks a specific format or structure, often found in text, images, and audio.

Valuable insights. Meaningful and actionable information extracted from data analysis, often used to make informed decisions.

Vector data. Vector data represents information that is transformed into numerical vectors, making it suitable for machine learning tasks.

Vectorizer. A tool or technique used to convert textual data into numerical representations for machine learning.

CHAPTER I

INTRODUCTION

Project Context

The rise of online shopping platforms began in the late 1990s and early 2000s; with the appearance of e-commerce and the ever-increasing accessibility of the Internet. Companies like Amazon, eBay, and Alibaba emerged as the pioneers in online shopping platforms, providing consumers with a convenient way of purchasing products online and having them delivered right to their doorsteps. Over the years, the popularity and usage of online shopping platforms have grown exponentially, with advancements in technology, secure payment systems, and improved logistics. The widespread adoption of smartphones and the development of mobile shopping applications further accelerated the use of online shopping platforms in the 2010s.

The COVID-19 pandemic has forced people to stay home, preventing them from shopping in physical stores. According to Morgan Stanley (2022), Global e-commerce increased from 15% of all retail sales in 2019 to 21% in 2021. This has led to a greater reliance on online shopping platforms to meet their essential needs and purchase necessary items. As a result, the adoption of online shopping platforms has experienced an unparalleled surge. Even in the post-pandemic era, still, a significant majority of individuals continue to prefer online shopping as their primary method of making purchases due to its convenience. The e-commerce market has a lot of room to expand in the long run and could go from $3.3 trillion in 2022 to $5.4 trillion in 2026. Consequently, online shopping platforms have become an integral part of the retail industry, with a growing number of consumers worldwide depending on them for their shopping needs.

In recent years, the rise of online shopping platforms has practically transformed how people shop, providing a convenient and accessible range of products. With the increasing popularity of e-commerce, customer satisfaction has become a crucial factor for the success of different online businesses. Understanding customer satisfaction levels and identifying factors influencing satisfaction can help businesses make data-driven decisions to improve their products, services, and overall customer experience.

Product reviews play a significant role in shaping customer perceptions and satisfaction. Customers often express their opinions, experiences, and feedback through reviews, providing valuable insights into their satisfaction levels. Analyzing these reviews can help businesses gain a deeper understanding of customer sentiment, identify areas for improvement, and make informed decisions to enhance customer satisfaction.

Purpose and Description

The prevalence of online shopping allowed many people to share their thoughts about a product through online reviews. These reviews serve as a valuable source of information for businesses. But with such an overwhelming number of reviews available, it becomes challenging for businesses to understand their customers' perspectives. The process of manually going through and comprehending the substantial volume of feedback is going to be overwhelming, resource-intensive, and time-consuming. There is a need for efficient and effective methods to extract meaningful insights from the sea of information.

While online shopping platforms provide a platform to express their opinion about a product, they are limited to the predominant reliance on simple rating scores. These scores, often represented by star ratings and numerical values, encapsulate an overall assessment but don't provide detailed information to understand the nuances of a customer's experience with the product. There is a need to do manual analysis to find what aspects the customers enjoyed or found lacking in the product. The lack of detailed information hinders the ability of businesses to tailor their products to meet customer expectations effectively. Without a deeper understanding of the reasons behind the customer's review, businesses have a limited understanding of how to refine their offerings. This limitation prevents small businesses from making targeted improvements and adjustments that could enhance customer satisfaction and loyalty.

Objectives of the Study

The primary goal of this study is to develop a system that can analyze customer reviews and extract key insights for decision support of small businesses to improve their offerings which in turn benefit the customers.

Specifically, this study aims to:

1. Develop a machine learning model that can classify customer satisfaction (positive, negative, neutral).
2. Streamline the process of reviewing multiple product reviews by allowing the user to provide a list for a more efficient input process.
3. Create a functionality that will give an overall analysis of customer satisfaction on a set of product reviews and provide the keywords or phrases that shaped the analysis.

Significance of the Study

Top Management. The proposed system equips top management with actionable insights, allowing them to make strategic adjustments to their policies that enhance products, refine marketing strategies, and improve customer support. Top managers can adapt to evolving customer preferences by making data-driven decisions.

Small Businesses. This study provides a potential for small businesses to drive growth and sustainability. Implementing customer feedback-driven changes increases the company's reputation, fosters loyalty, and boosts profitability. The effectiveness of the policies will enable these businesses to make their own brand identity, thus standing out in a crowded market.

Customers. The ripple effect of these improvements in the businesses will extend to the customer. Customers benefit from tailored products and services, exceeding their expectations. Enhanced satisfaction, trust, and loyalty result from businesses that actively listen and adapt, creating a more rewarding purchasing experience.

Researchers. This study serves as a foundation of knowledge for the researchers. It lets them develop and refine their computer science skills, such as data preprocessing and machine learning. The practical application of these techniques can further enrich their research capabilities.

Future Researchers. Future researchers can benefit from the study by building upon its findings and methodology. The study guides and inspires for conducting similar research or exploring related topics.

Scope and Limitation

The choice of a machine learning algorithm for this study is the Support Vector Machine (SVM) which is known for its effectiveness in classification tasks. The SVM algorithm is utilized for training a model that categorizes the reviews as positive, neutral, or negative based on extracted features; The study considers English Language product reviews from a global perspective. As a result, the model's performance and accuracy might not generalize well to reviews in other languages. The inclusion of multilingual capabilities would require additional data collection, preprocessing, and potentially adapting the model, which falls beyond the scope of our current investigation. Therefore, the analysis and predictions provided by our model may be limited in their applicability to non-English reviews; The datasets that will be utilized to train the machine learning model are pre-existing. The contents of the datasets are taken from the online shopping platform Amazon which houses a vast number of English product reviews on a variety of products.

The web application operates independently, meaning it is designed as a standalone system that does not integrate or connect with other software programs. This decision was made to streamline development and focus on the core functionalities of the sentiment analysis. While the term “standalone” might benefit from clarification, it essentially signifies that the developed web application functions autonomously. Additionally, the application is specifically tailored as a web-based platform, and we have opted not to extend its development to include a software application (whether as desktop or mobile). Although this addition may seem to balance the scope, it aligns with the proponent’s decision to concentrate on web application development.

There are also limitations that the study would like to specify. The model’s analysis will be based solely on the textual content of product reviews. It will not take into account other forms of feedback such as images, emoticons, videos, or audio reviews; The model may have limitations in capturing the full complexity of customer opinions, as sentiment analysis might not accurately capture subtle emotions or sarcasm. The study assumes that the extracted features accurately represent the sentiments expressed in the reviews, which may not always be the case due to human language's inherent complexity and variability. The model might also lean on commonly used words and won't have a perfect analysis of slang or uncommon words further highlighting the challenges posed by linguistic complexity; Another one of the limitations of this study is its reliance on the quality of the dataset. Star ratings as the basis for the positive, negative, and neutral customer reviews and thus relies on the dataset's label correctness. The dataset might also contain flaws such as inconsistencies or inarticulateness of the reviewer. To address these potential issues, data preparation will be done, such as the removal of duplicates and attempts to fill in missing information. While these steps aim to enhance the reliability of the dataset, we acknowledge that some degree of label dependency and quality variability may persist.

The web application system is constrained by a limit on the number of reviews it can effectively analyze. This limitation is deliberately imposed to align with the research-oriented nature of our project, allowing us to manage the volume of data within the scope of our study. It's important to note that grammar-checking functionality is intentionally excluded from the system's capabilities and falls beyond the defined scope. This strategic decision was made to maintain the focus on sentiment analysis, acknowledging that grammar checking is a distinct and complex domain that requires a dedicated approach.

CHAPTER II

RELATED LITERATURE

The related literature section comprises foreign and local literature that serves as a reference for the present study. It delves into the interaction between online shopping, consumer feedback, and machine learning algorithms, providing valuable insights and knowledge that contribute to the investigation for this thesis study.

Related Literature

Foreign Literature

There has been extensive discourse on analyzing customer satisfaction concerning online shopping platforms. Perara (2023) emphasizes the crucial role of customer reviews in gauging satisfaction levels. The author emphasizes how sentiment analysis, utilizing natural language processing techniques, can yield valuable insights from product reviews as they provide valuable feedback to potential customers and help build trust in a brand or product. Furthermore, in the article about why customer sentiment analysis matters by Patel (2023) also emphasizes the significance of sentiment analysis in online retail and its impact on customer satisfaction. These two collectively highlight the importance of developing an effective machine learning model to assess product reviews and ascertain customer satisfaction within online shopping platforms.

Machine learning algorithms have been widely discussed for classifying product reviews into positive, neutral, or negative sentiments. An article by Wolff (2020) about sentiment analysis and machine learning explores various machine learning techniques, including Support Vector Machine, for sentiment analysis. The author discusses SVM's efficacy in classifying reviews and its potential application in the e-commerce domain. Additionally, an online article about solving an important problem of detecting suspicious reviewers in online discussions on social networks by Machova, Mach, and Vasilko (2021) compares different machine learning algorithms for sentiment analysis, and emphasizes the accuracy and efficiency of SVM in categorizing product reviews. These sources highlight SVM as a promising algorithm to classify product reviews, providing valuable insights into customer satisfaction in online shopping platforms.

As Mhatre (2022) highlighted, Python stands out as a versatile and contemporary general-purpose programming language renowned for its amalgamation of power, speed, and accessibility. Python's execution through interpreters ensures cross-platform compatibility, making it adaptable across various operating systems. Its extensive application spectrum encompasses web platforms, graphical interfaces, data science, and machine learning. Beklemysheva (2020) states that Python provides concise and easily comprehensible code. While complex algorithms and versatile workflows stand behind machine learning and AI, Python’s simplicity allows developers to write reliable systems. Developers get to put all their effort into solving an ML problem instead of focusing on the technical nuances of the language. Python code is understandable by humans, which makes it easier to build models for machine learning. Notably, Python has seen a surge in adoption within the realm of data analysis and holds a prominent position as one of the most widely utilized languages in the field of data science.

Local Literature

The creation of this study stemmed from the proponents' interest in the growing popularity of e-commerce. In the article by Chan (2023) about statistics and trends of different Filipino online shopping behavior, e-commerce global retail sales increased from 13.6% to 18% during the pandemic. As a result, the competition among businesses has intensified, necessitating a deeper understanding of Filipino online shopping behavior to gain strategic advantages. Filipinos spend around 10 hours daily on social media, representing a significant majority demonstrating higher spending and active engagement, making it a powerful platform for advertising and selling products. Furthermore, the Philippines experienced the highest online shopping growth in Southeast Asia during the pandemic, and the projected sales are expected to reach $24 billion in 2025. Reyes (2021) in an article also states that sentiment analysis during the pandemic can be a critical tool to easily provide guidance and influence decisions of policy-makers to address the pressing issues.

In the context of Pinoy businesses, Jermaine Delos Santos wrote an article based on a study in 2023 by GoDaddy. It discovered that 51% of Filipino respondents were startups or entrepreneurs who established their commercial ventures within one to five years, slightly higher than the global average of 46%. Notably, Filipino businesses prioritize establishing an online presence, with 49% creating social media accounts and 30% engaging in online advertising. This approach has proven fruitful, as 62% of participants attribute up to half of their annual revenues to online sales channels, and 38% credit over half or even the entirety of their income to online operations. The success of this strategy is further supported by the Philippines' 67% rating for having a strong online presence, surpassing the global average of 58%.

In the column by Ordinario (2021), inferred that among the ASEAN countries covered in the study, the Philippines emerged as the most reliant on ratings, reviews, and opinions of friends and family. Based on a report by iKala on Southeast Asia Social Commerce, it was found that 86% of Filipinos depend on ratings and reviews when making online purchases. The study further revealed that 44% of Filipino consumers rely on official company information from websites, while 35% look to influencers for guidance. About 33% also value customer support in their decision-making process. Only 5% of Filipinos do not gather information through social media before purchasing. Crismundo (2022) further confirms this in his report about how most pinoy e-shoppers read reviews first, which details a study conducted in the same year by Ninja Van and DPD group on Southeast Asia E-commerce. The researchers of the study claimed that the majority of Filipino online buyers first check product reviews before making a decision. Filipinos value reviews to such an extent that they actively share their feedback with family and friends or post it on e-commerce platforms. Additionally, Filipino online shoppers recognize the significance of reviews in avoiding difficulties with returning defective products or wrong orders.

Related Studies

The collection of studies contains international and domestic research that acts as a reference for the current study. It examines the intertwining influence of online shopping, consumer feedback, and machine learning algorithms.

Foreign Studies

In a research paper about sentiment analysis for online reviews using conditional random fields and Support Vector Machines by Xia et al. (2020), they explored the realm of sentiment analysis concerning online product reviews. The study highlighted the significance of extracting customer sentiments from textual data to gauge their satisfaction levels. By employing sentiment analysis techniques, they revealed how understanding customer satisfaction patterns and sentiments expressed in reviews can be crucial. Sentiment analysis, as an automatic method for discerning sentiments in online reviews, holds considerable importance and can find applications in diverse fields beyond customer feedback analysis. Additionally, Vijayaragavan, Ponnusamy, and Aramudhan (2020), with their study about an optimal support vector machine-based classification model for sentiment analysis of online product reviews, highlights the significance of using natural language processing (NLP) techniques and machine learning algorithms to analyze sentiment in product reviews. NLP aids in understanding customer preferences expressed in unstructured data, allowing developers to identify positive or negative sentiments toward product features. By considering negation and tracking sentiments across time, designers can quantitatively assess the impact of design extensions and make data-driven decisions for product improvement. This application of NLP enables businesses to gain valuable insights from customer feedback, leading to enhanced customer satisfaction and product development. Both research studies emphasize the importance of developing a trustworthy and resilient machine learning model capable of efficiently analyzing product reviews. Such a model can offer valuable customer satisfaction insights, crucial in the rapidly changing world of online shopping platforms.

According to Diekson et al. (2022), machine learning algorithms like Support Vector Machine often use numerical data, and the proponents will need to transform or convert the data into a set of numerical vector data with a process commonly known as vectorization. In their study about a case study of Traveloka for sentiment analysis of customer reviews, they mentioned that this vectorizer will convert input data by calculating how much the TF-IDF score for each word in the dataset and finally put the information into a vector form. In a study about sentiment analysis on e-commerce platforms reviews using SVM algorithm by Hossain, Joy, Das, and Mustafa (2022) that TF-IDF is one of the accurate extraction technique to solve the challenge of the machine learning cannot understand strings. TF-IDF converts strings to numbers in order to provide numerical data to machine learning models. Also stated by Irawaty, Andreswari, and Pramesti (2020) in their case study about comparison of vectorizer for sentiment analysis on YouTube, TF-IDF is the best vectorizer in this research because according to the results the study come up with is that there are no errors in predicting negative values, and also has many positive predictive values compared to other vectorizers. In relation to the proponent's study, it is best to do the classification using SVM algorithm with TF-IDF Vectorizer.

When exploring the machine learning algorithms suitable for classifying product reviews into positive, neutral, or negative sentiments, various research studies have been conducted to investigate their performance and accuracy. In a comparative study by Nguyen, Veluchamy, Diop, and Iqbal (2018), the authors evaluated multiple algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), and Gradient Boosting, to determine the most effective approach for sentiment analysis. Their research findings indicated that SVM exhibited superior performance in distinguishing sentiment categories, showcasing its potential as a robust classifier. Moreover, in the study conducted by Hossain, Joy, Das, and Mustafa (2022), they employed SVM to classify sentiment in e-commerce reviews. Their results demonstrated that SVM achieved high accuracy in predicting sentiments, reinforcing its suitability for sentiment analysis in the context of product reviews. With these studies, it becomes evident that the SVM algorithm offers an accurate and reliable approach to classify product reviews based on different sentiment classes effectively.

Prema, Hema, Maheshprabhu, and Nageswara (2021) used Linear SVC in their study about sentimental analysis to get an idea about what people in Twitter have in their minds or get people's emotions using Machine Learning Algorithms. Linear

Support Vector Classification is a module of Scikit Learn. SVM and LinearSVC are related concepts, where Support Vector Machine is the general algorithm used for classification and regression, and Linear SVC is a specific implementation of SVM for linearly separable data. Linear SVC tends to fit the provided data and returns the 'Best Fit' hyperplane, through the fit method through their studied sample_weight parameter. According to their statistics, Linear SVC got a 98.8314% accuracy in training their data, and a 77.5233% accuracy in testing data, which is the best accuracy among 6 other classification performances of their tested algorithm. In a study about machine learning based automated disaster message classification system that used Linear SVC algorithm by Prasanna et al. (2023), they stated that Linear SVC (Support Vector Classifier) is one of the most successful linear models, not only because it is quite fast to train and compute, but also because it can achieve excellent performance in high dimensional problems. Linear SVMs can make use of a wide variety of learning algorithms. Their work uses Linear SVC Algorithm with the strategy of self-training that learns from available datasets with the labeled data.

When TF-IDF combined with the Support Vector Machine algorithm, according to Arthamevia, Adiwijaya, and Purbolaksono (2021) in their study of an aspect-based sentiment analysis using TF-IDF and SVM algorithm, they obtained an accuracy of 91.24%. Thus, they concluded that the sentiment classification process in this research was carried out more accurately by using SVM algorithm. They tried different algorithm to go with TF-IDF like Naive Bayes, Logistic Regression, Decision Tree, and Ensemble Methods, and was proven that SVM algorithm to be the higher-ranking above them. Based on the experimental results in the study about four kernel functions on SVM algorithm with TF-IDF by Prastyo, Ardiyanto, and Hidayat (2020), the Linear kernel function on SVM+TF-IDF using 4000 features has achieved the best performance of all feature tests in accuracy, precision, recall, and f-measure with the value 96.53%, 96.61%, 96.50%, 96.52%, respectively. In addition, TF-IDF can increase SVM's performance using 1000 until 4000 features. According to their study, Linear SVC with TF-IDF gives one of the highest accuracies of getting the accurate classification. TF-IDF is employed as a feature extraction and selection to improve SVM performance.

In their research, Nguyen, Veluchamy, Diop, and Iqbal (2018) explored how different methods can classify the sentiment of the text. They looked at six approaches: three using machine learning - Logistic Regression, SVM, and Gradient Boosting, and three using lexicons - Valence Aware Dictionary (VADER) and Sentiment Reasoner, Pattern, and SentiWordNet. They tested these approaches on Amazon reviews, and used accuracy, precision, recall, and F1-score for performance measures. The results showed that the machine learning models performed better than the lexicon-based which were the bottom 3. The accuracy results are as follows: LR (90%), SVM (89%), Gradient Boosting (87%), VADER (83%), SentiWordNet (80%), Pattern (69%). The study suggests that machine learning works better than lexicon-based approaches.

The proponents investigated the suitable software development life cycle (SDLC) model for their research. In a conference paper, Laato et al. (2022) discussed the integration of machine learning with software development lifecycles. According to the data scientists they interviewed, the waterfall model was favored due to its straightforward nature and readability (p. 9). However, the researchers pointed out that the traditional waterfall model is not very good at returning to earlier stages of development. To address this issue, they suggested looking into versions of the waterfall model that include iterative elements (p. 3).

Local Studies

Goyenechea, Nartea, and Santos (2019) identified different factors affecting Filipino Millennials' shopping behavior in their study. One of the factors considered is feedback, which has a Moderate Positive Correlation. While Filipino Millennials mainly agree with relying on the online store's reputation. The respondents also generally agree that positive online reviews influence their choice of online stores and that they also rely on other customers' reviews of the online store's service. Their agreement indicates that positive reviews play a role in influencing their choice of online stores, and they are more likely to prefer stores that have favorable feedback. The researchers also emphasized that people that read online reviews about products before making a purchase might also recommend the same compared to those that do not.

Since product reviews are seen as one of the significant factors in purchasing online, it is sensible to analyze whether consumers are satisfied with their purchase. Because manual categorization for big datasets can be burdensome, supervised machine learning techniques can be used to train models to analyze Customer Satisfaction based on Natural language. Corpuz (2021) tried to compare two machine learning algorithms to categorize customer satisfaction. The algorithms used were SVM and Long Short-Term Memory (LTSM) Neural Networks. The results show that both SVM and LTSM have high training accuracy, with SVM scoring 98.63% and LSTM scoring 99.32%. While they perform similarly in accuracy, SVM is significantly faster, taking about 35.47 seconds less.

There have also been attempts to analyze Product Reviews using Hybrid Approach. Ceniza-Canillo, Montejo, and Sagarino (2022) created a study about sentiment analysis on product reviews which utilized VADER for annotation and two machine learning models hence the hybrid approach. The two models were Multinomial Naive Bayes (MNB) and SVM. The final results indicate that positive reviews comprise 83.6% of the total, negative reviews constitute 9.1%, and neutral reviews make up 7.3%. The SVM model outperformed MNB with an 83% accuracy score. The researchers also conducted a survey to validate the mode's results, and it confirmed that 75.8% of respondents would recommend a store or product based on positive reviews.

Feliscuzo, Gorro, and Romana (2022) conducted research that focused on improving the accuracy of the SVM classifier for text data by using data word vectorization techniques. The techniques explored were TF-IDF, CountVectorizer, and Word2vec skip-gram with custom weight using TF-IDF. The initial results showed that the TF-IDF-based model achieved an accuracy of 80%, and the CountVectorizer-based model achieved an accuracy of 63%. Meanwhile, the combination of Word2vec skip-gram and TF-IDF as the custom weight resulted in an accuracy of 86%. Based on the results, the researchers concluded that a hybrid approach could improve the accuracy of the SVM classifier given the text data.

In a recent study by Ignacio, Sicangco, and Esquive (2023), the researchers successfully trained a machine learning model and deployed it on a web application to access the model easily. Their approach employed the iterative waterfall model, enabling them to jump through the development stages when making and evaluating multiple models. The study highlights the success of their chosen approach in achieving favorable model results.

Synthesis

The study originates from the surge in e-commerce, with global online retail sales growing to 18% from 13.6% during the pandemic, as Chan (2023) explains. The Philippines, a key player, experienced the highest online shopping growth in Southeast Asia, expecting $24 billion in sales by 2025. Filipinos' heavy social media use—around 10 hours daily—drives engagement and spending, making it a potent platform for product promotion. Delos Santos (2023) insights further confirm this trend, highlighting that Filipino businesses are prioritizing their online presence, leading to substantial revenues from online channels. With a strong online presence rating of 67%, surpassing the global average, businesses in the Philippines are embracing this digital shift. This scenario underscores the study's focus on product reviews' role in customer satisfaction, aligning well with the booming e-commerce landscape in the country. Reyes (2021) also underscores the potential significance of sentiment analysis during the pandemic, serving as a valuable tool to guide policymakers and address pressing issues.

Ordinario's findings (2021) highlight the Philippines' distinct reliance on product ratings and reviews within ASEAN countries, with 86% of Filipino consumers depending on these factors for online purchases, as revealed by iKala's report. Similarly, Crismundo's report (2022) reinforces this trend, indicating that most Filipino e-shoppers prioritize reading reviews before making decisions. Goyenechea, Nartea, and Santos (2019) add that Filipino Millennials' shopping behavior is influenced by positive online reviews, demonstrating a moderate positive correlation. The tendency to recommend products and stores based on positive reviews is further underscored by survey validation results, where 75.8% of respondents expressed willingness to do so. These insights collectively illuminate the profound impact of product reviews on customer satisfaction, emphasizing their critical role in shaping consumers' choices and fostering trust in businesses, an aspect crucial to the study's exploration of customer satisfaction analysis in the realm of online shopping platforms.

The discourse on customer satisfaction analysis in online shopping emphasizes sentiment analysis and Support Vector Machine (SVM) as crucial tools. Perara (2023), Patel (2023), Wolff (2020), Machova et al. (2021), Xia et al. (2020), and Vijayaragavan et al. (2020) collectively highlight SVM's accuracy and efficiency in classifying sentiments, crucial for extracting insights from product reviews. In this study's context, SVM proves a robust approach for customer satisfaction analysis, aligning with the goal of enhancing understanding of user sentiments in online shopping platforms. The proponent’s thesis aims to leverage SVM and sentiment analysis, contributing valuable insights to businesses through real-time customer satisfaction analysis.

In the context of this study focused on customer satisfaction analysis of product reviews in online shopping platforms, the utilization of Linear Support Vector Classification (Linear SVC) as a kernel for Support Vector Machine (SVM) gains significance from related research. Prema, Hema, Maheshprabhu, and Nageswara (2021) demonstrated the effectiveness of Linear SVC in sentiment analysis, emphasizing its role in capturing emotions and opinions from social media data. Linear SVC, being a specific implementation of SVM for linearly separable data, is designed to determine the 'Best Fit' hyperplane for classification tasks. Its remarkable accuracy of 98.8314% in training and 77.5233% in testing, as observed in their study, underscores its robust performance. Additionally, Prasanna et al. (2023) highlighted Linear SVC's success as a rapid yet powerful linear model, particularly suitable for high-dimensional problems. The algorithm's ability to excel in learning from labeled datasets, as well as its efficacy in automated disaster message classification, further attests to its potential in delivering accurate and efficient results. Thus, in line with this study's objective, the selection of Linear SVC as the SVM kernel resonates with the algorithm's proven track record of achieving exceptional performance, ensuring reliable sentiment classification of online product reviews.

In the realm of sentiment analysis for product reviews within online shopping platforms, various studies have been conducted to ascertain the efficacy of different machine learning algorithms. Nguyen, Veluchamy, Diop, and Iqbal (2018) undertook a comparative study that evaluated multiple algorithms, encompassing Logistic Regression (LR), Support Vector Machine (SVM), and Gradient Boosting.

Their findings highlighted SVM's notable performance in sentiment classification, positioning it as a robust classifier. Additionally, Hossain, Joy, Das, and Mustafa (2022) utilized SVM to categorize sentiment in e-commerce reviews, revealing its high accuracy in predicting sentiments. Notably, Corpuz (2021) compared SVM and Long Short-Term Memory (LSTM) Neural Networks for customer satisfaction categorization, with both exhibiting high training accuracy. SVM's swift computational prowess was particularly emphasized, setting it apart from LSTM. Another intriguing avenue involved a hybrid approach by Ceniza-Canillo, Montejo, and Sagarino (2022), who employed VADER annotation in conjunction with SVM and Multinomial Naive Bayes (MNB) models. Here, SVM emerged as the frontrunner with superior accuracy, bolstered by survey validation indicating a strong inclination toward recommendations based on positive reviews. Collectively, these related literatures and studies underscore SVM's consistent accuracy and suitability, positioning it as a highly effective choice for the proponents to use it in their study.

In relation to the proponent’s study of customer satisfaction analysis in online shopping, combining Support Vector Machine (SVM) with TF-IDF vectorization proves effective. Diekson et al. (2022) highlight SVM's reliance on numerical data, echoed by Hossain et al. (2022) and Irawaty et al. (2020) regarding TF-IDF's accuracy in converting text to numbers for machine learning. Arthamevia et al. (2021) showcase SVM-TF-IDF's 91.24% accuracy, while Prastyo et al. (2020) demonstrate the superiority of Linear SVM with TF-IDF. Feliscuzo et al. (2022)

support TF-IDF's consistent accuracy. These findings support SVM-TF-IDF's suitability for accurate sentiment analysis of online shopping reviews.

Technical Background

Customer satisfaction analysis involves the analysis of customer feedback on products. This includes product ratings, product reviews to gauge their level of satisfaction.

The proponents will develop a system that utilizes data science and machine learning techniques to study data and extract meaning from it. With the use of natural language processing to understand, interpret, and generate human language. This includes data collection, data preprocessing, feature extraction and data visualization of the product reviews. Machine learning involves detecting patterns or regularities within data. Analyzing the data collected can help understand the underlying process or make discoveries on the patterns that can be used to create predictions (Alpaydin, 2020). With the use of machine learning algorithms, we can devise a model that can be automated to analyze the customer feedback and create predictions.

Python serves as the primary programming language used in this study, seamlessly integrating all the different components of the system. Python is known for its power, speed, and accessibility (Mhatre, 2022). Beklemysheva (2020) emphasizes Python's ability to simplify the coding of the machine learning model, allowing developers to focus on solving complex problems like sentiment analysis. Python 3.11.4 is the latest (as of writing) and chosen version for this study. The reason for the selection is that Python 3 is the current standard, and this version is designed to address the limitations and security issues of the previous versions (Koubbi, 2022; St John, 2023). Python's dominance in data analysis further supports its suitability for efficiently analyzing customer reviews in online shopping, which aligns with the proponent's research goals.

For this research, the proponents employed the Support Vector Machine algorithm (SVM) as the primary classifier because of its ability to handle data where there are features to consider. This is well-suited for the task of separating data into different classes. The model will be using a pre-existing dataset in order to learn from labeled data and create its own understanding from it. The dataset will be split into multiple subsets to use one part for testing and the rest for training.

In the context of training the model, the proponents will first use a web-based Integrated Development Environment (IDE) as the testing environment to find the essential parameters that yield the best SVM model. When the parameters are identified, the model will be saved as a file and incorporated with the other functionalities of the system using the code editor.

CHAPTER III

RESEARCH METHODOLOGY

The nature of the study can be classified as developmental research, primarily because it entails the systematic design, creation, and deployment of a system. By following this approach, the study systematically crafts a decision support system. This involves the creation of a machine learning model and a web application to streamline the customer satisfaction analysis product reviews in online shopping platforms using Support Vector Machine.

Software Design

When looking for a suitable SDLC, the proponents turn to the findings of existing research papers. Laato et al. (2022) emphasized the waterfall model's potential for integrating machine learning into software development, suggesting iterative enhancements. Meanwhile, Ignacio et al. (2023) employed the iterative waterfall model to train a machine learning model and successfully deployed it on a web application. Using these findings, the researchers find that the model aligns with the study by utilizing its iterative components to enhance adaptability and development efficiency across the phases.

Thus, the proponents opt to use the Modified Waterfall Development due to the iterative process of the machine learning project. The complexity of machine learning and natural language processing tasks requires a structured and systematic development process. The flexibility of the Modified Waterfall Model to refine the system as challenges emerge makes it a good fit for the system.

This model, also known as the Iterative Waterfall Model, is a variant of the traditional waterfall model which breaks down the software development process into phases that are completed sequentially. Feedback paths are included in the Modified Waterfall model to allow the iterations between phases.

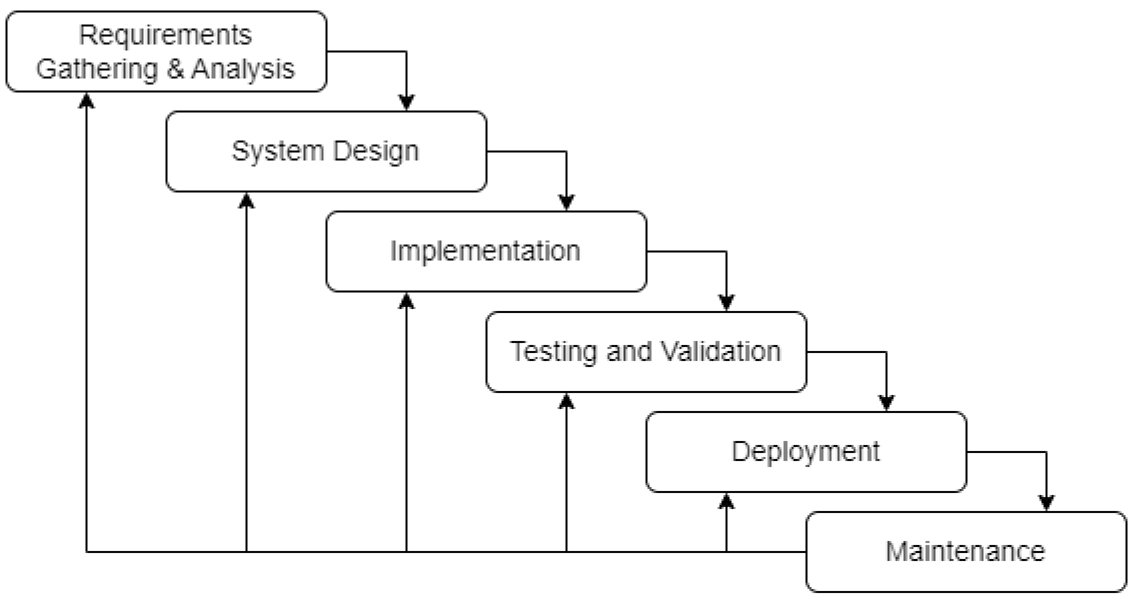

Figure 1. Modified Waterfall Model

Requirements Gathering

In the initial phase, the proponents engage in discussions, consultations, and research to outline the requirements for the study. The research problem and objectives are defined in order to give purpose and goal, and the scope and limitations of the study are also identified. Additionally, the requirements for the system functionalities are documented.

The researchers notice that the abundance of online product reviews is challenging to analyze manually. Current reviews lack detailed information beyond rating scores thus obscuring the customer feedback. As an answer, the primary focus of the research is to develop a system capable of classifying customer satisfaction, streamlining the user input, and extracting meaningful insights from product reviews.

System Design

The system's flow is designed in this phase following the requirements gathered from the previous phase. The proponents decide to employ SVM as the machine learning algorithm to classify the feedback expressed due to its effectiveness in text classification tasks. The design also includes the development of a web application that lets the user provide a single review or a set of reviews for more efficient output. The system is designed to analyze the reviews, extract keywords or phrases, and perform an individual and overall customer satisfaction analysis.

Implementation

The implementation phase involves the actual development of the system. The research team writes the code to create the machine learning model using the SVM algorithm for sentiment classification. Python is chosen as the primary programming language for its extensive libraries for natural language processing and machine learning. The user interface is developed using web development technologies to access the model easily.

Testing and Validation

Thorough testing and validation are essential to ensure the accuracy and reliability of the system. Amazon product reviews are used to train and test the machine learning model. The iterative process allows for tuning models. This method allows for the creation of multiple models with different parameters.

The models' performance is evaluated using precision, recall, F1 score, and accuracy metrics. This evaluation allows an overview of each model's strengths and areas that require improvement.

Deployment

After successful testing and validation, the best model is selected and implemented in the system. The system is then deployed as a web application for users to interact with.

Maintenance

Troubleshooting is done to fix bugs and errors that arise from the system. Refinements to the machine learning model, user interface, and overall functionalities are also done to improve the system's effectiveness and usability. The modified waterfall design allows the iterative development of the machine learning model.

Conceptual Design

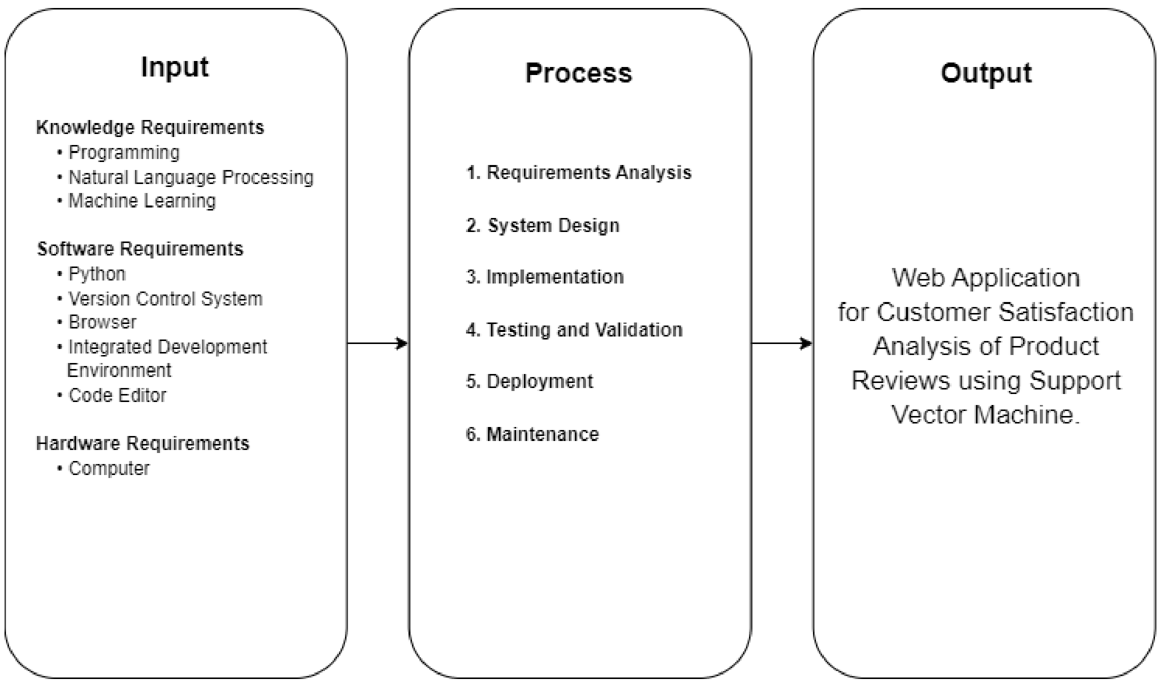

Figure 2. The Conceptual Framework

Figure 2 outlines the study's Input, Process, and Output Model, which serves as the guide as the research development occurs. The model starts with the Input Stage, which includes the fundamental requirements gathered during the requirements gathering. This stage is divided into three parts; The Knowledge Requirement specifies the skills needed to develop the machine learning model and the web application system. Meanwhile, the Software Requirement defines the technologies needed to build the system, such as Python, Version Control System, Browser, Integrated Development Environment, and Code editor. Lastly, The Hardware Requirement is the computer.

The next stage is the process which includes the actual development of the system, which includes the requirements analysis, system design, implementation, testing and validation, deployment, and maintenance. This phase follows the SDLC as stated in the Software design.

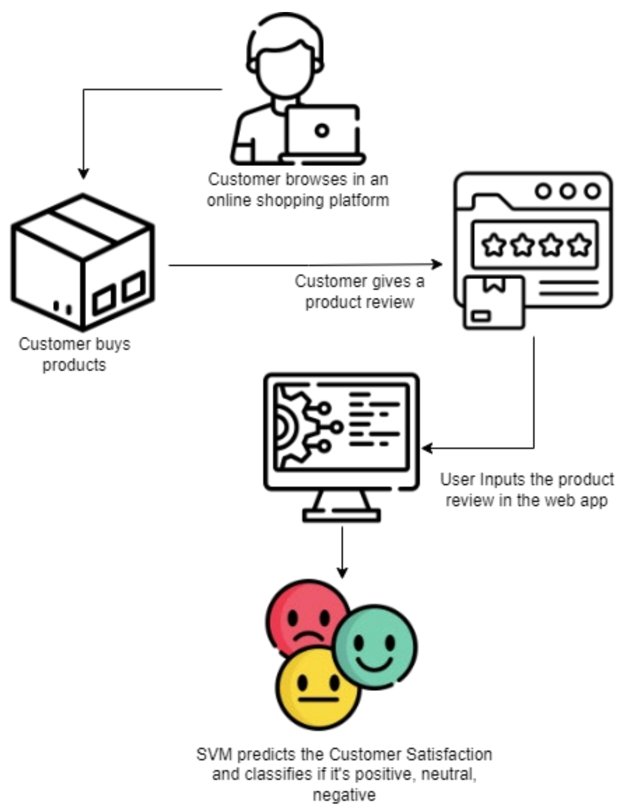

Figure 3. Simplified Use Case Diagram

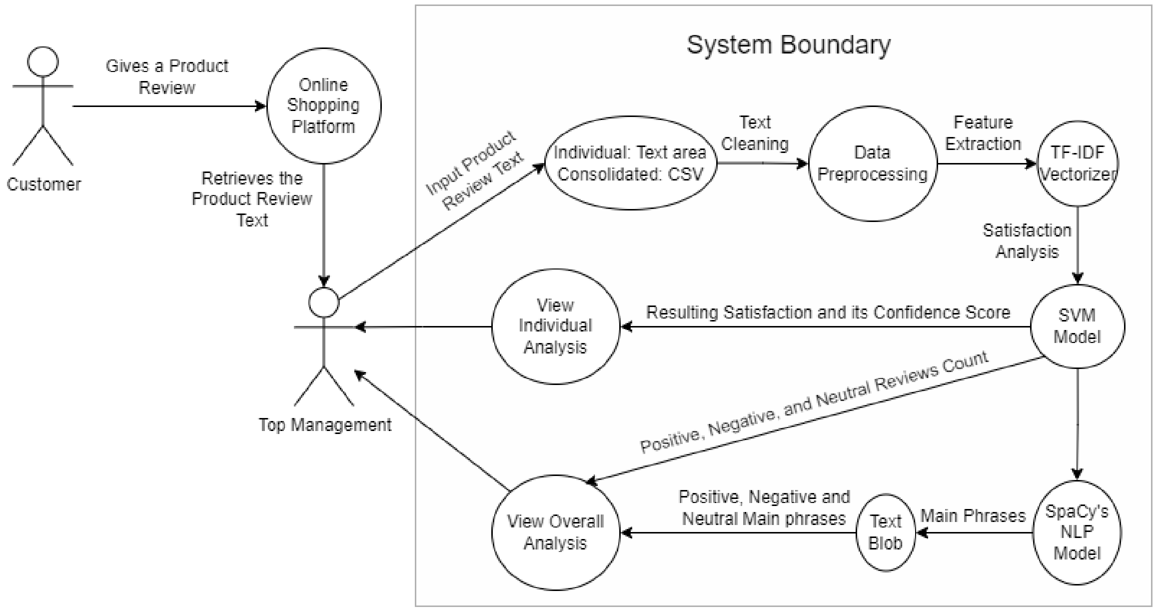

Figure 4. Detailed Use Case Diagram

The proponents aim to assess customer satisfaction with online shopping experiences by analyzing product reviews on various online shopping platforms. The goal is to determine whether these product reviews are positive, neutral, or negative and extract key insights that allow the proponents to gauge customer satisfaction levels with the products and services they have availed. To efficiently analyze a number of product reviews, the proponents employ a machine learning algorithm called the Support Vector Machine (SVM). This tool quickly helps to categorize the product reviews as positive, neutral, or negative, enabling us to process the feedback faster than manual methods.

In Figure 4, the primary user engaging with the system is the "Top Management." Their interaction begins with inputting review text through the "Input Review" interface, where the user will input a single review for individual analysis and a CSV list of reviews for consolidated. Once confirmed, the reviews will go through data preprocessing for text cleaning, and the processed data travels to the "Support Vector Machine" for sentiment analysis. Subsequently, the results are further processed for consolidated analysis.

For individual reviews, the "Analyze Sentiment" command directs the processed data to "View Individual Analysis," presenting the result of the individual review, whether it is positive, negative, or neutral, and its corresponding confidence score.

On the other hand, for consolidated reviews, the "Analyze Sentiment" command directs the processed data to "Extract Insights," where meaningful keywords or phrases are identified. This outcome is consolidated in "View Overall Analysis," comprising sentiment analysis outcomes and extracted insights presenting a comprehensive overview of Customer Satisfaction to the "Top Management." To elaborate the data processing of consolidated analysis in the SVM model, sentiment occurrences are tracked using variables such as "positive_count," "negative_count," and "neutral_count." Simultaneously, the reviews undergo Natural Language Processing (NLP) using Spacy to extract phrases based on patterns like "ADJ," "NOUN," and "ADV," along with punctuation marks. These phrases are stored in a list called "review_phrases." The list of phrases then proceeds to another sentiment model (TextBlob), which predicts sentiment and categorizes phrases into three lists: "positive_phrases," "negative_phrases," and "neutral_phrases." Upon completion, the system provides a summary, including the total review count, sentiment counts (positive, negative, neutral), and lists of significant phrases for positive, negative, and neutral sentiments. These results are then displayed in the 'result_page.html.'

By leveraging Support Vector Machine (SVM), the proponents gain a comprehensive understanding of customer preferences and concerns regarding the products. This insight is valuable for businesses as it helps them identify areas of strength and areas that require improvement. Understanding customer satisfaction is crucial for businesses to maintain a happy customer base and attract new customers.

Application Requirement

The system's development relies on various software and hardware components working together harmoniously to implement the customer satisfaction analysis. **Python 3.11.4**, the selected version for this study, serves as the primary programming language, offering the latest features and functionalities at the time of research.

Software Requirements:

- **Web-based Integrated Development Environment** – serves as the testing environment to find the parameters for the best machine learning model.

    - **Local – Jupyter Notebook:** Ideal for data science tasks and machine learning algorithm development, providing a flexible and user-friendly interface.
    - **Cloud – Google Colab:** A cloud-based version of Jupyter Notebook, utilized for researchers without sufficient computing power by allowing 12.7 GB of RAM.

- **Code Editor – VScode:** Used to polish the system's code, integrating the model with the other functionalities of the system.
- **Version Control System – Git:** Used to track changes and manage code revisions, ensuring seamless collaboration among researchers and version control for the project.
- **Github:** A web-based hosting service for Git repositories, serving as a centralized platform for storing and sharing the codebase and related documentation with the research team.
- **Browser:** A browser is necessary to access the web application of the machine learning model, allowing users to interact with and utilize the model's functionalities through a graphical user interface.

Hardware Requirements:

The following hardware is the specification of the computer used to develop the system:

- **Processor:** AMD Ryzen 5 3600

- **Storage:** Samsung 970 EVO PLUS SSD
- **RAM:** 16 GB
- **Power Supply:** MSI MPG 750 watts
- **GPU:** Nvidia RTX 3060 ti

System Architecture/System Flow

The system flow illustrates the progression of building a machine learning model, encompassing steps from dataset preparation to model training and evaluation. It culminates in the deployment of the Support Vector Machine (SVM) model within a web application for real-time customer satisfaction analysis.

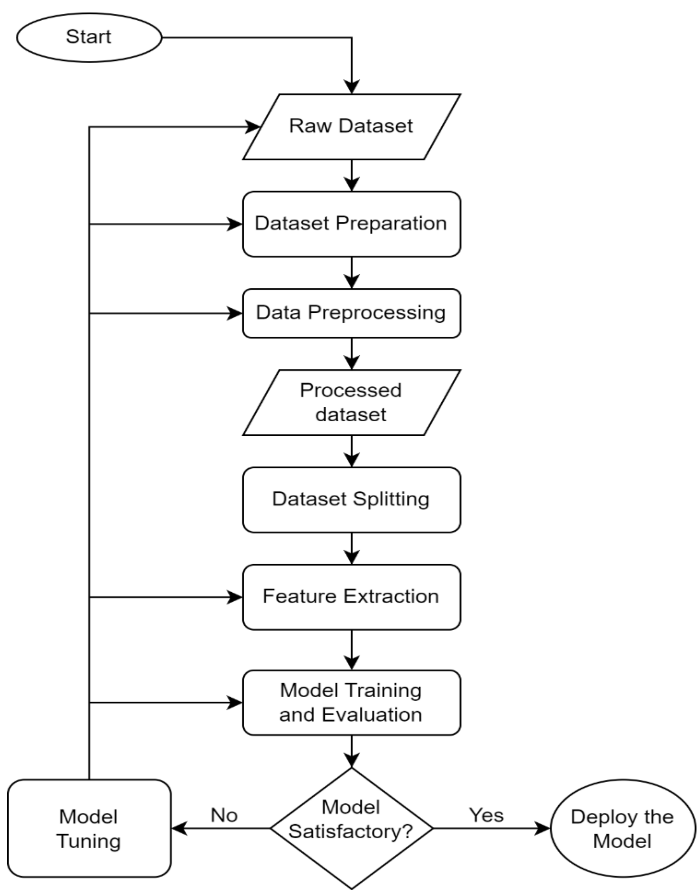

Figure 5. System flow of machine learning model training

1. **Dataset Preparation**

   a. **Dataset Loading.** The first step involves loading the dataset containing product reviews into the system. This dataset will be the basis for training and testing the machine learning algorithm.

   b. **Field Selection.** From the loaded dataset, select the relevant fields or attributes that are essential for the analysis. These fields may include the product review text and the product ratings or score, which are relevant information.

   c. **Missing and Duplicate Data Checks.** Before proceeding, the system checks for any missing or duplicated data points in the dataset. If found, appropriate actions are taken, such as data imputation or removal of duplicates.

   d. **Feedback Mapping of Scores.** To perform sentiment analysis, map the numerical ratings (scores) in the dataset to corresponding sentiment labels, such as positive, neutral, or negative.

   e. **Resampling Techniques.** In cases where the dataset is imbalanced (i.e., one sentiment class dominates over others), apply resampling techniques to balance the representation of different sentiments. This ensures a more robust model training.

2. **Data Preprocessing**

    a. **Text Cleaning:** The system undergoes text cleaning to remove unnecessary characters like symbols and HTML tags, ensuring only meaningful text remains. Normalization, such as making the text lowercase, is also applied in order to standardize the formatting and make it more uniform.

    b. **Expanding the Contraction of Texts:** Texts with contractions are expanded to their full form to improve the consistency and understanding of the text data.

    c. **Negation Handling:** The system identifies negations in the text (e.g., "not happy") and handles them appropriately to avoid misinterpretation of sentiments.

    d. **Stop Word Removal:** Commonly used words that do not add significant meaning to the text (e.g., "the," "is," "and") are removed to reduce noise and improve processing efficiency.

3. **Dataset Splitting**

After getting the vectorized values, the dataset is split into two separate sets: the training set and the test set. The training set is used to train the machine learning model, while the test set is used to evaluate its performance on unseen data.

4. Feature Extraction

The processed text data is transformed into numerical representations suitable for machine learning tasks. Vectorization techniques convert textual data into numerical representations, enabling efficient mathematical operations essential for training the machine learning algorithm. This transformation allows the algorithm to process and analyze text-based information effectively, leading to accurate feedback classification and customer satisfaction analysis.

The processed text data undergoes a transformation into numerical representations suited for machine learning tasks, using methods like TF-IDF (Term Frequency-Inverse Document Frequency), which turns words into numbers.TF-IDF facilitates the conversion of text data into numerical representations, thereby heightening the efficiency of training algorithms, notably within the domains of natural language processing and information retrieval.

TF(Term Frequency) Formula

$$TF = \frac{\text{Number of repeat words in sentence}}{\text{Number of words in sentence}}$$

IDF(Inverse Document Frequency) Formula

$$IDF = \frac{\text{Number of Sentence in Dataset}}{\text{Number of Sentence in dataset that contains specific word}}$$

TFIDF Formula

$$TFIDF = TF * IDF$$

Figure 6. TF-IDF Formula

The TF-IDF vectorizer applies the formula converting a set of text documents into a matrix of TF-IDF features. It takes into account both the frequency of a term in a document (TF) and the inverse document frequency (IDF), which penalizes terms that are too common across all documents.

5. Support Vector Machine (SVM) Model Training:

The system trains an SVM classifier using preprocessed and vectorized training data. The SVM model learns from the labeled data to classify reviews into their respective sentiment categories (positive, neutral, or negative).

The study focuses on Linear as the kernel due to its simplicity, computational efficiency, and compatibility with linearly separable datasets. By utilizing the Linear SVM kernel, the study aims to achieve precise feedback classification while maintaining an easily interpretable model.

6. Model Testing and Evaluation

The SVM model is then tested on a separate set of preprocessed and vectorized test data to assess its performance on new, unseen reviews.

After training the SVM model, it is evaluated on a separate set of preprocessed and vectorized test data to assess its performance on new, unseen reviews. The evaluation of sklearn.metrics include the following:

a. **Accuracy Score:** The accuracy score measures the percentage of correctly classified reviews out of the total reviews in the test set. It provides an overall assessment of the model's performance.

b. **Confusion Matrix:** The confusion matrix presents a tabular summary of the model's predicted classifications against the actual sentiments. It shows the number of true positives, true negatives, false positives, and false negatives, helping to evaluate the model's ability to correctly classify each sentiment.

c. **Classification Report:** The classification report provides a comprehensive evaluation of the model's precision, recall, and F1-score for each feedback class (positive, neutral, and negative). Precision measures the accuracy of positive predictions, recall calculates the percentage of correctly identified positive cases, and the F1-score balances precision and recall.

7. Model Tuning

If the current model’s evaluation is not satisfactory, revisions of the data and model might be needed. Model tuning is the iterative process of enhancing the performance and quality of a machine learning model by making systematic adjustments to various components of the model.

8. Model Deployment in Web App

Finally, the trained SVM model is integrated into the web application to enable real-time customer satisfaction analysis. Users can input their product reviews through the user-friendly interface, and the SVM model provides analysis results promptly, offering valuable insights into customer satisfaction levels

Input and Output Reports and Analysis

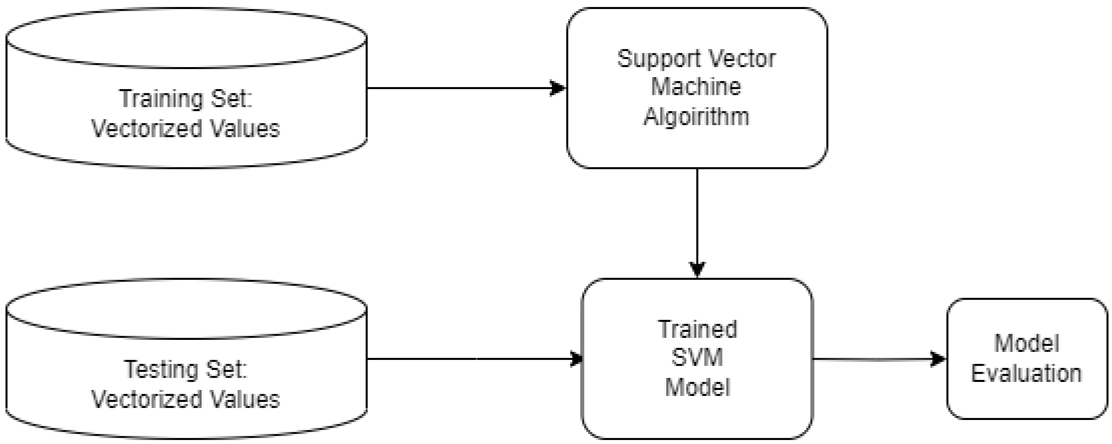

Figure 7. Dataset Input to Model Evaluation Output

After going through text mining with text cleaning and preprocessing, the processed data is split into two categories— Training and Testing data. Training data teaches a model to recognize patterns and make predictions. This data comprises product reviews and annotated feedback labels (positive, neutral, negative). On the other hand, testing data contains new and unseen reviews to assess if the model can correctly classify them.

Both data undergo feature extraction to transform textual data into vectorized values that the algorithm can understand and process. These numerical representations capture essential information about the reviews, allowing the Support Vector Machine (SVM) to analyze customer feedback and determine their satisfaction effectively. The SVM algorithm learns and becomes a trained SVM model. The model is validated using the testing data, and the evaluation results into different measurements representing how well the model did during the test.

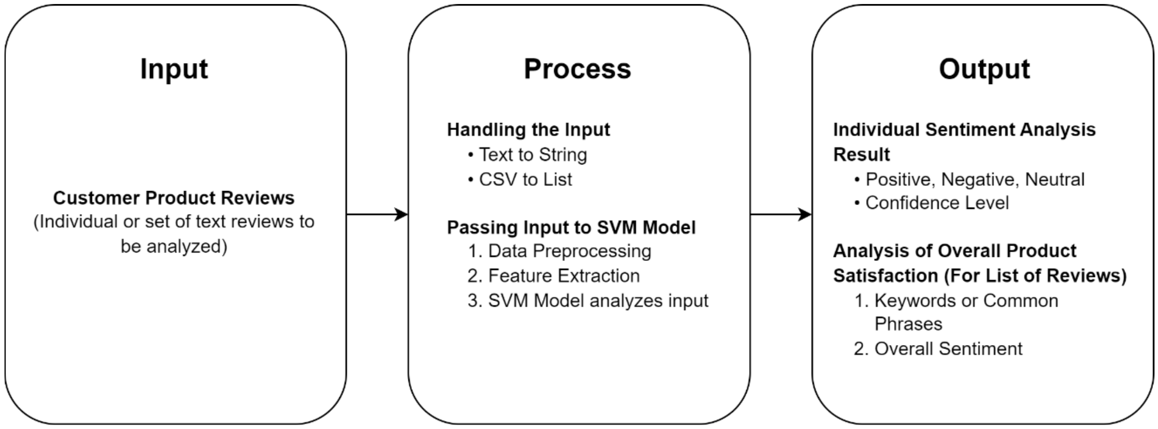

Figure 8. User Input to Classifier Output

The system consists of an input, and the output is accessible via the web application while an additional process is happening in the background.

The first step is for the users to provide the customer's product reviews in the system. The reviews are typically text, it can be placed in the input box or have the set of reviews in csv and imported into the web application. When the user triggers the button to confirm their input, it will be sent to the back end.

The input data first goes through the text mining process. It includes tasks like text cleaning (normalization and removing unnecessary symbols, HTML tags, and links) and Text pre-processing (expanding contractions, handling negations, and removing stop words). Text mining is necessary since this ensures that the user input has less noise and is more consistent. This step makes the data more neat and optimized. After that, feature extraction converts the unstructured data (text) into a structured format that a machine learning algorithm can understand and process. The structured format is typically numerical representations in which algorithms can easily find patterns and relationships. After the conversion, it is fed into the trained support vector machine model, which analyzes the input and then makes its inference. The model makes a classification, whether positive, neutral, or negative, and is sent back to the user as an output.

Multiple Constraints

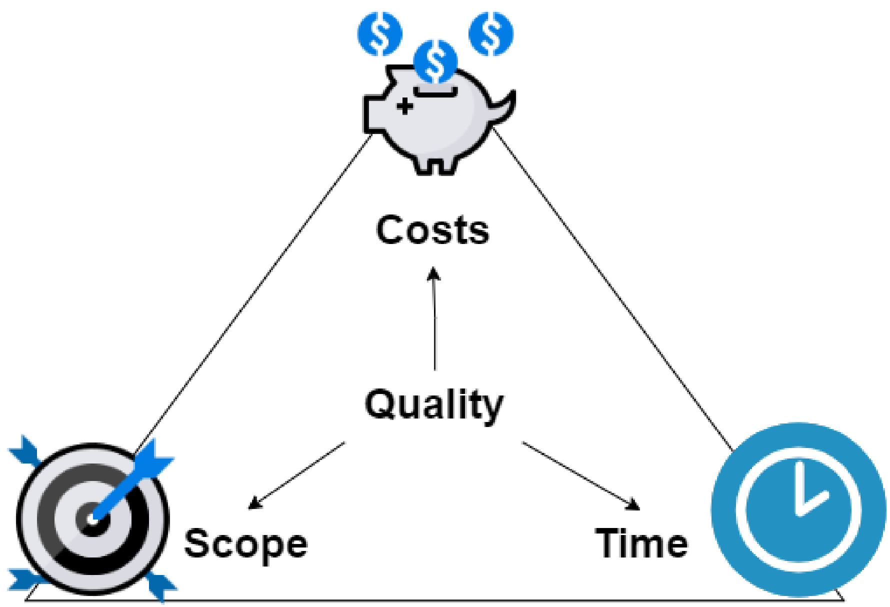

Figure 9. Visualization of Project Constraints

The proposed study's scope is narrowed down to using a single machine learning algorithm for analyzing customer sentiment and English text restriction. This decision is made to maintain a focused scope within the available resources and timeline. Implementing multiple algorithms would require more resources and time for experimentation and validation. And the proponents recognize there are different languages available, and accommodating them would introduce additional complexities in the data collection and model development. Focusing on a single algorithm and language allows for a better model. This approach aligns with the study’s objective of providing a meaningful decision support system for small businesses while working within practical limitations.

While the proposed study benefits from the availability of a free dataset and open source software, the proponents acknowledge that other non-financial costs still apply. The absence of direct monetary expenses may offset the financial burden, but the project still incurs opportunity costs in terms of time and effort. Data preprocessing, model development, and analysis require dedicated resources from the proponents. These non-monetary costs should not be underestimated, as they impact the project's overall feasibility. The commitment of time and effort could limit the depth of analysis or the extent of system refinement, especially considering the study's multi-faceted objectives.

Algorithm Use

Introduction to Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm widely used for classification tasks. In this study, Support Vector Machine (SVM) is chosen for its effectiveness in feedback analysis and customer satisfaction analysis based on product reviews. Support Vector Machine (SVM) aims to find the optimal hyperplane that best separates different feedback classes, allowing us to categorize product reviews as positive, neutral, or negative accurately.

Mathematical Formulation of SVM

In the context of customer satisfaction analysis, let's represent the input data or the text data as a set of feature vectors x, where each vector corresponds to a product review. Additionally, let y be the feedback class label, which can take on values 1 (positive), 0 (neutral), or -1 (negative). The Support Vector Machine (SVM) seeks to find a hyperplane represented by the equation $w \cdot x + b = 0$ that best separates the positive and negative feedback classes. The margin, represented as $Margin$, is the distance between the hyperplane and the closest data points, known as support vectors. Support Vector Machine (SVM) aims to maximize this margin, as it leads to better generalization and improved classification accuracy.

Hyperplane Equation and Decision Boundary

The hyperplane equation $w \cdot x + b = 0$ allows SVM to make decisions regarding the feedback class of a given product review. For instance, if $w \cdot x + b > 0$, the review is classified as positive, and if $w \cdot x + b < 0$, it is classified as negative. Reviews falling on the hyperplane ($w \cdot x + b = 0$) are considered neutral. The decision boundary is formed by the hyperplane, and SVM aims to find the hyperplane that best separates the classes with the maximum margin.

Linear SVC

Linear SVC is a SVM algorithm for linearly separable data, aiming to find the best hyperplane to separate classes while maximizing margin. Originally designed for binary classification, it can be extended to multiclass problems using various strategies.

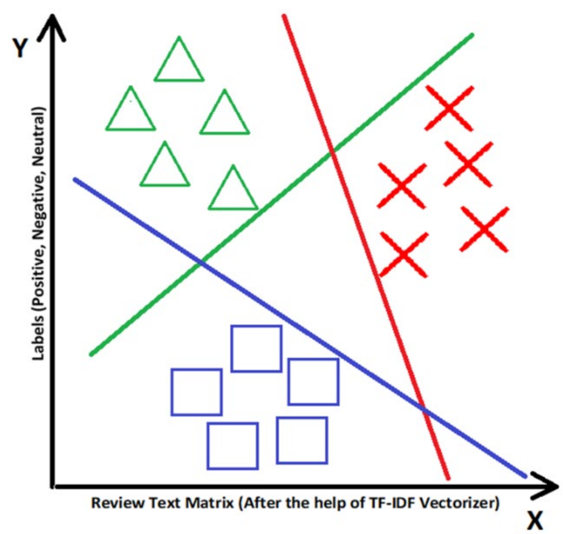

Figure 10. LinearSVC for Multi Classification

After the TF-IDF vectorization process transforms review text into a structured matrix, the resulting enriched TF-IDF matrix captures the essence of each review by encapsulating individual terms and their contextual relationships. This matrix is subsequently integrated into the Linear Support Vector Classification (Linear SVC) algorithm, allowing it to leverage the comprehensive information distilled from the reviews for the purpose of sentiment classification.

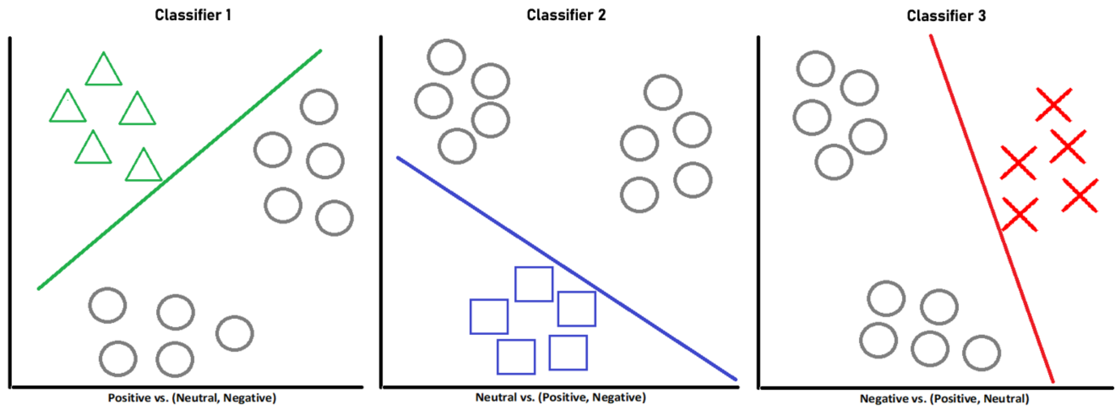

Figure 11. LinearSVC Hyperplane in Multi Classification using OVR Approach

The Linear SVC algorithm, a variant of SVM, is extended to accommodate multi class scenarios using the one-vs-rest (OvR) strategy.

One-vs-Rest (OvR) Approach

Classifier 1: Positive vs. (Neutral, Negative)

Classifier 2: Neutral vs. (Positive, Negative)

Classifier 3: Negative vs. (Positive, Neutral)

The one-vs-rest (OvR) method extends the Linear SVC technique, a form of SVM, for multiclass situations. In order to forecast the input review sentiment, it compares transformed features to the decision limits of OvR binary classifiers and uses the highest confidence score.

CHAPTER IV

RESULTS AND DISCUSSION

A. Model Training and Evaluation

The dataset used to create the model comprises reviews from various categories, specifically sourced from the Amazon Review Data by Jianmo Ni (2018). The dataset selection process involved downloading different categories and extracting relevant columns, namely "Text" and "Score." To ensure diversity, a maximum of 10,000 rows per category were selected, resulting in a dataset containing 9,999 reviews per category. It's important to note that digital goods reviews were excluded, and the possibility of incorporating them as a limitation needs consideration.

In crafting our sentiment analysis model for online shopping websites, a crucial decision was made to strategically sample a specific percentage from the available dataset instead of using the entire corpus. This choice was motivated by the dual objectives of optimizing training efficiency and mitigating resource constraints while ensuring the selection of a representative dataset subset.

The decision to sample a specific percentage was validated through a comprehensive model evaluation process. The model underwent training on the sampled dataset, and the results exhibited a satisfactory level of accuracy and predictive performance. This process not only affirmed the effectiveness of our chosen sampling strategy but also underscored the broader importance of having distinct training and testing datasets.

Training the model on a subset of the data and evaluating it on a separate, unseen subset is a fundamental practice in machine learning (Mitchell, 1997). This approach helps gauge the model's ability to generalize to new, unseen data. The selected percentage for the training dataset is critical, as it directly impacts the model's capacity to learn patterns and relationships within the data.

File Sizes		
Dataset	**Vectorizer**	**Model**
10%	47.37 MB	72.46 MB
20%	82.19 MB	126.51 MB
30%	112.51 MB	173.75 MB
40%	140.44 MB	217.38 MB
50%	166.52 MB	258.21 MB
60%	190.32 MB	295.53 MB
70%	213.39 MB	331.76 MB
80%	234.98 MB	365.68 MB
90%	256.45 MB	399.47 MB
100%	276.27 MB	430.68 MB

Table 1. Equivalent File Sizes per Dataset Usage Percentage

Model Evaluation	
Dataset	**Accuracy**
10%	87.755 %
20%	88.100 %
30%	88.396 %
40%	88.229 %
50%	88.412 %
60%	88.524 %
70%	88.552 %
80%	88.575 %
90%	88.552 %
100%	88.613 %

Table 2. Model Evaluation per Dataset Percentage

Table 1 provides insights into the file sizes associated with different dataset percentages, emphasizing the need for a careful balance between data size and model efficiency. Additionally, Table 2 breaks down model evaluations, focusing on accuracy as a key metric for various dataset percentages. These tabulated comparisons facilitate the identification of the most suitable dataset percentage for achieving optimal file sizes and predictive accuracy (Bishop, 2006).

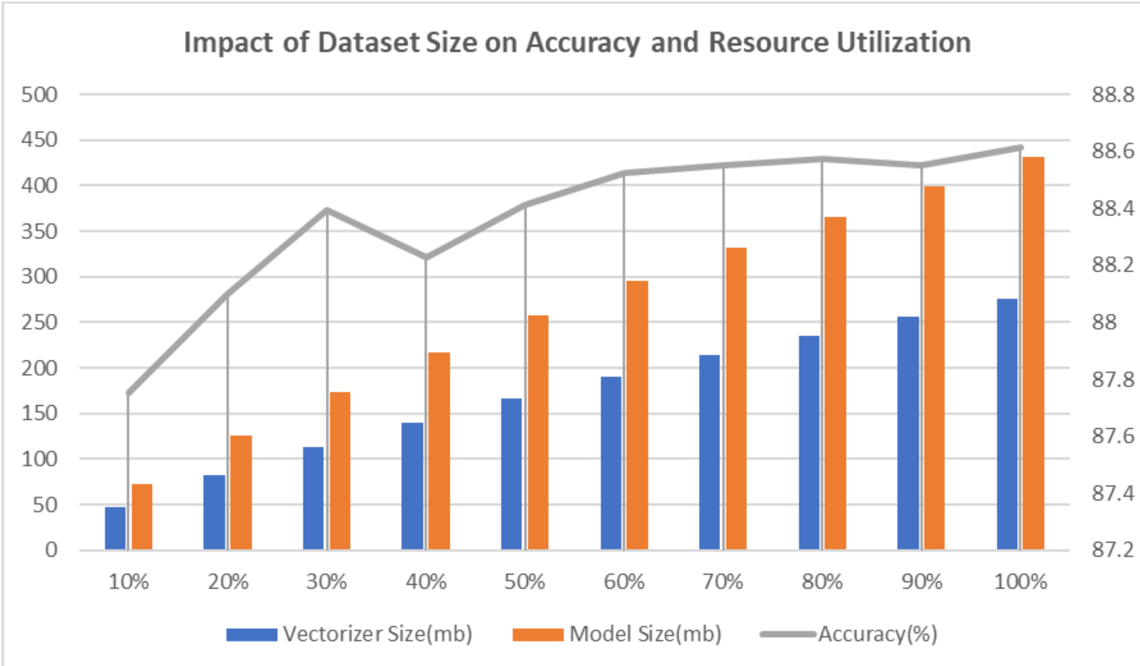

Figure 12. Impact of Dataset Size on Accuracy and Resource Utilization

The selection of the testing data percentage is a crucial decision in model training and evaluation, balancing accuracy with resource efficiency. The figure above shows and illustrates the trade-off between accuracy and file size for different testing data percentages.

As observed, the accuracy generally increases with a larger dataset percentage, reaching its peak at 100%. However, a significant increase in file size accompanies this rise, indicating a potential trade-off in terms of computational resources and storage requirements. The substantial growth in file size might lead to diminishing returns, where the marginal increase in accuracy does not justify the considerable expansion in resource usage.

Therefore, 30% of the dataset was selected as it achieves a commendable accuracy of 88.396% while striking a balance in its file sizes. This suggests that the model performs optimally and offers a favorable compromise between accuracy and resource efficiency. Furthermore, observations show that the dataset accuracy exhibited a dip at 40% and a slight dip at 90%. This dip could be attributed to increased noise or variability introduced by the additional data, leading to a minor reduction in predictive performance. The marginal decline in accuracy and the significant increase in file size makes selecting 30% of the dataset a strategic choice.

In conclusion, selecting 30% for the testing dataset represents a pragmatic compromise, delivering a high level of accuracy while managing computational resources effectively. This decision ensures that the model is robustly tested without incurring excessive computational costs associated with larger datasets.

Another method to evaluate the model is through the use of test cases. By incorporating test cases that encompass various scenarios, such as the usage of emoticons, special characters, capital letters, or numbers, the model is subjected to real-world conditions where user-generated content can vary widely. Handling these scenarios effectively is vital so that the program can continue to process and handle errors accordingly.

Testing the model against these varied test cases is akin to a stress test, examining its performance under different content structures. A robust sentiment analysis model should exhibit flexibility in handling diverse inputs, making it adept at extracting sentiments regardless of the presence of noisy and unstructured data.

This comprehensive evaluation ensures that the model's capabilities extend beyond a narrow set of linguistic patterns, enhancing its applicability to a wide range of real-world scenarios.

By distinctly partitioning the datasets and using test cases to check if functioning properly, our methodology adheres to the foundational principles of machine learning, fostering a robust sentiment analysis model poised for effective real-world applications. The ensuing section will delve into the nuanced details of this model training and evaluation process, elucidating the justifications behind the sampling decision and the resultant model performance.

B. Model Verification

To verify the model's performance, a novel approach is proposed: testing it on product reviews from different online shopping platforms like Flipkart and eBay. The process involves reconfiguring the data to match the model's specifications. The objective is to evaluate how well the model generalizes to datasets beyond its training source (Amazon). This approach introduces a practical application by testing the model's predictive capabilities on real-world data from diverse sources.

Here's a breakdown of the accuracy achieved on external datasets:

- eBay Dataset Accuracy: 93.82%
- Flipkart Dataset Accuracy: 84.19%
- Sephora Dataset Accuracy: 88.48%
- Shopee Dataset Accuracy: 90.99%

In assessing the model trained on the Amazon dataset's predictive accuracy on different platforms, compelling results were obtained. The model demonstrated a high level of accuracy across diverse datasets, including eBay (93.82%), Flipkart (84.19%), Sephora (88.48%), and Shopee (90.99%). This variance in accuracy prompts a closer look into potential influencing factors.

For instance, the analysis acknowledges that since the model was primarily trained on the Amazon dataset, its performance might be influenced by the unique characteristics of Amazon reviews. When applied to platforms like Flipkart, the language nuances, particularly considering that English may not be the first language of the primary user base, could impact accuracy. Another factor is the category; Sephora's focus on beauty and personal care products, while a strength for its niche market, could be a downside for the model’s prediction ability as it is trained on a more general category of product reviews.

Meanwhile, Ebay and Shopee consist of multiple English-speaking countries and have various categories of products, so this could be the reason why the model has done well in these datasets. Despite these differences, the model's overall performance is commendable, aligning with existing research that considers accuracy rates of 80% and above in sentiment analysis as indicative of a robust model, as emphasized by Smith and Jones (2018).

Zhang et al. (2017) highlighted in their paper the challenges of understanding generalization in deep learning models, emphasizing that a diverse range of product categories aids in better model generalization. They argue that training sentiment analysis models on reviews from various domains enable them to learn patterns and sentiments applicable across different industries, ultimately enhancing the model's ability to analyze sentiments more broadly. Moreover, Cambria and White (2014) provide insights into natural language processing research, indicating that understanding sentiments across industries is crucial. Aggarwal and Zhai's book on "Mining Text Data" (2012) supports this notion by covering various aspects of text mining, including sentiment analysis, and can provide valuable insights into its applications. Rokach and Maimon's handbook on "Clustering Methods" (2015) is relevant in the context of obtaining comprehensive customer satisfaction insights by exploring diverse patterns. Additionally, Bolukbasi et al. (2016) discuss methods to debias word embeddings, underlining the importance of avoiding biases in language corpora, especially when dealing with reviews from various domains in sentiment analysis. These references collectively stress the significance of considering diverse product categories for robust sentiment analysis in e-commerce platforms.

C. Developed Web Application

The RevU's is a website developed to analyze customer reviews and extract key insights for decision support of small businesses to improve their offerings, which in turn benefit the customers. The web application comprises Home, Features, Context, The Team, and the Analyzer, which consists of an Individual Review Analyzer and a Consolidated Reviews Analyzer. The pages are designed for top management and small businesses where they can easily input collated product reviews in the Individual and/or Consolidated Review Analyzer without manually analyzing the reviews. After submission of the Individual or Consolidated product reviews, the analyzer will provide the results (for Individual, the results will be on the page itself; for Consolidated, the results will be in the form of a pop-up.

Our web application embodies a user-centric design with a visually engaging landing page, where Sentivm, our mascot, dynamically represents various satisfaction levels, providing an immediate connection with users. The mascot's animations and mobile responsiveness enhance the overall user experience.

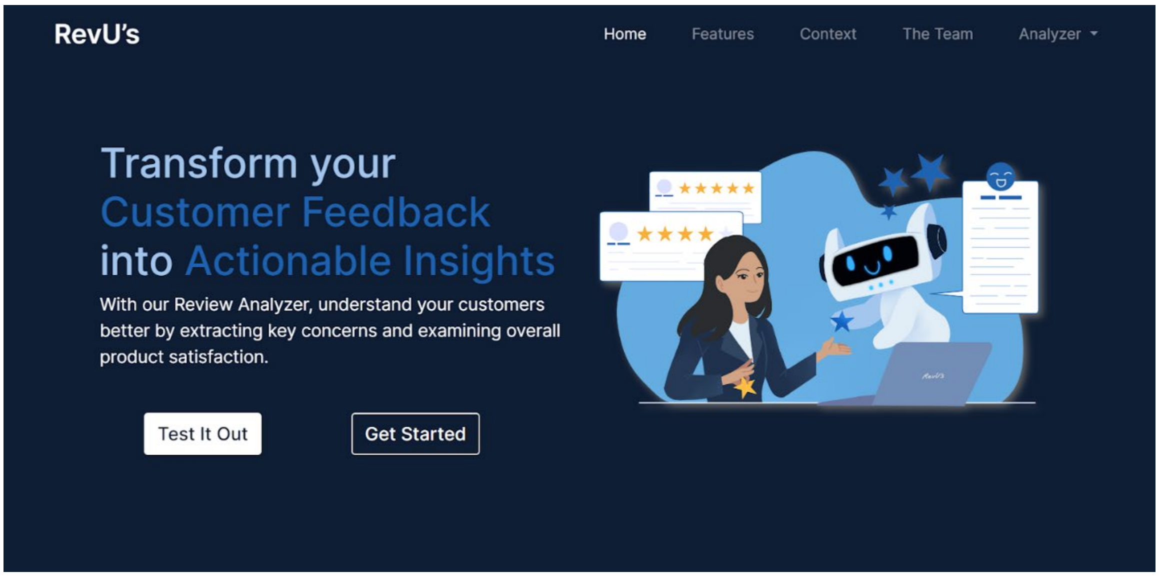

Figure 13. Home Page

The home page serves as the entry point, featuring Sentivm to create an inviting environment. Its purpose is to attract users, offering a brief intuitive introduction to the sentiment analysis journey.

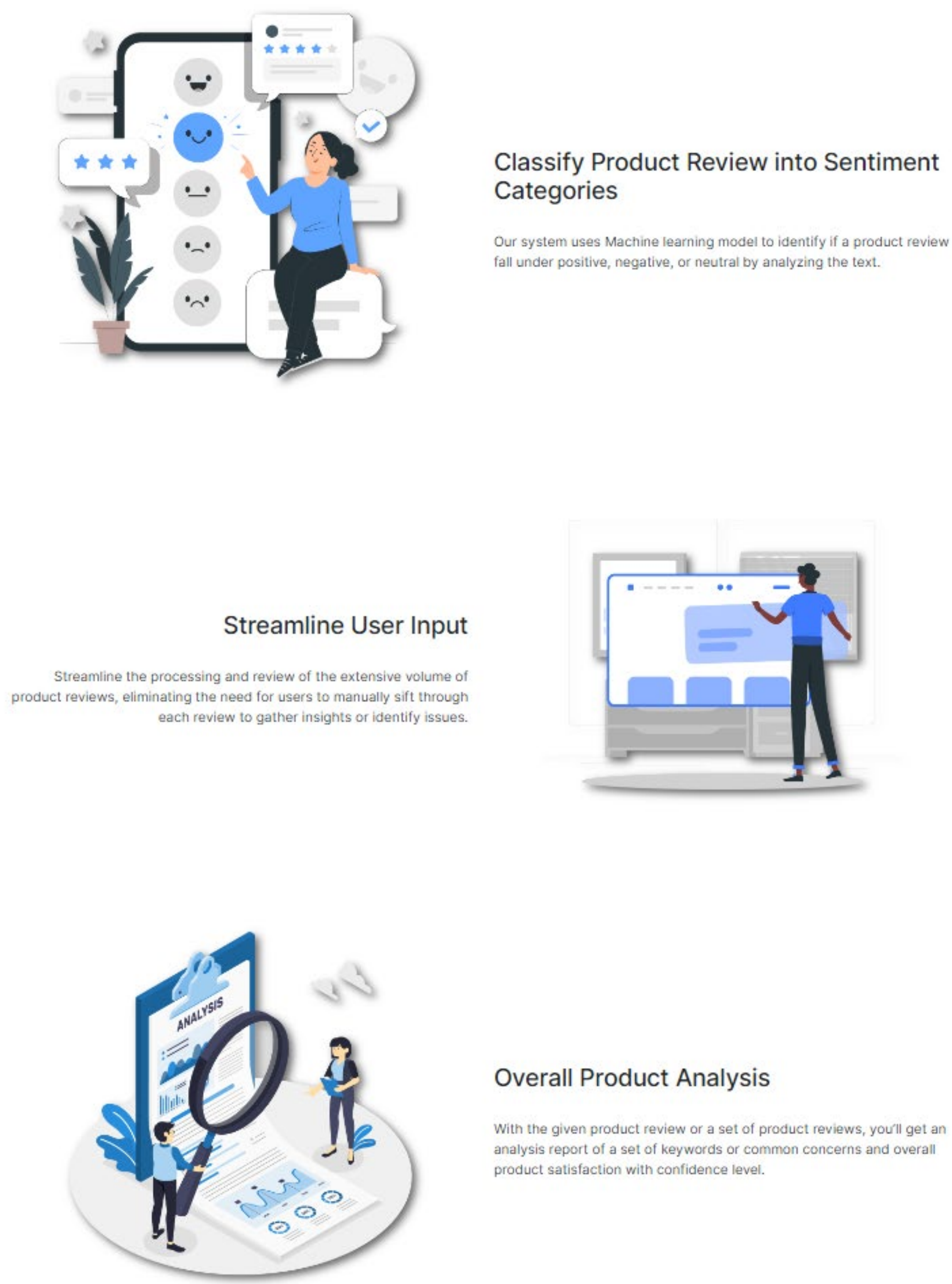

Figure 14. Features Page

The Features Page consists of what will be the outcome or the process when you use the RevU's Product Analyzer, whether it be the Individual or the Consolidated Analyzer, it shows you the pros and the process to be made to analyze your Product Reviews.

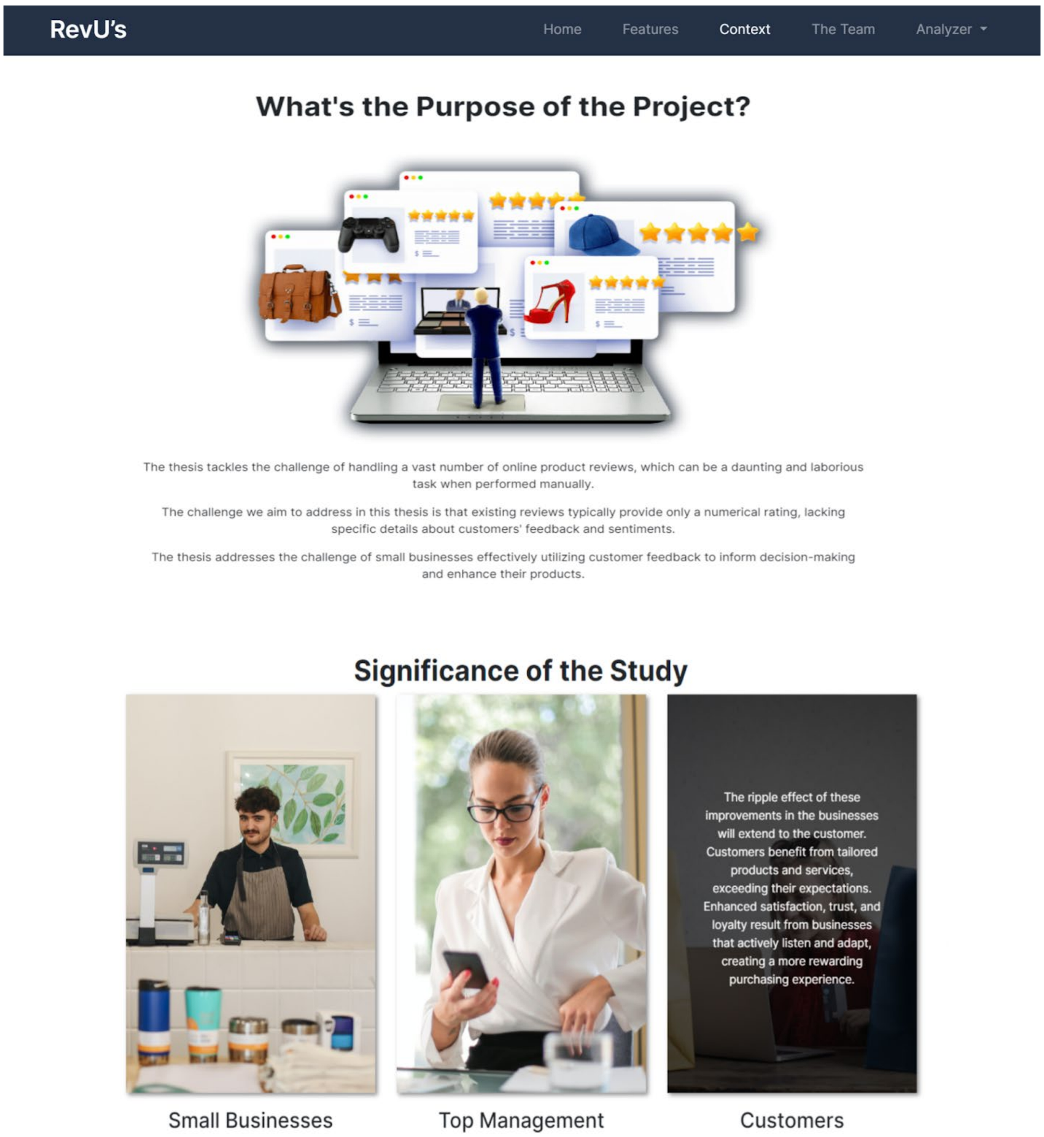

Figure 15. Context Page

The context page shows the purpose of creating this web application system as the proponent's thesis project. This page tackles the what and why's behind the idea of the RevU's web application. This is also where the proponents highlight the main individuals or groups that may benefit from the developed web application. Whether one will use this for personal use, or they'll use it to enhance their customer experience and enhance their products with the results.

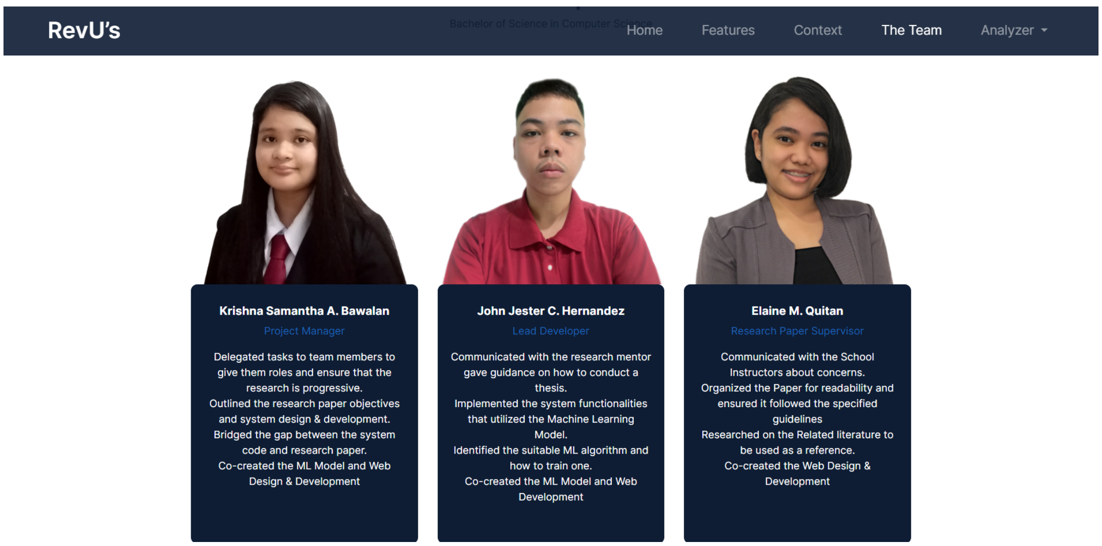

Figure 16. Meet the Team Page

A personalized touch is added through "The Team" page, introducing the creators behind RevU's. This is an additional way to let the users know who the creators are and what their roles are while creating the web application system.

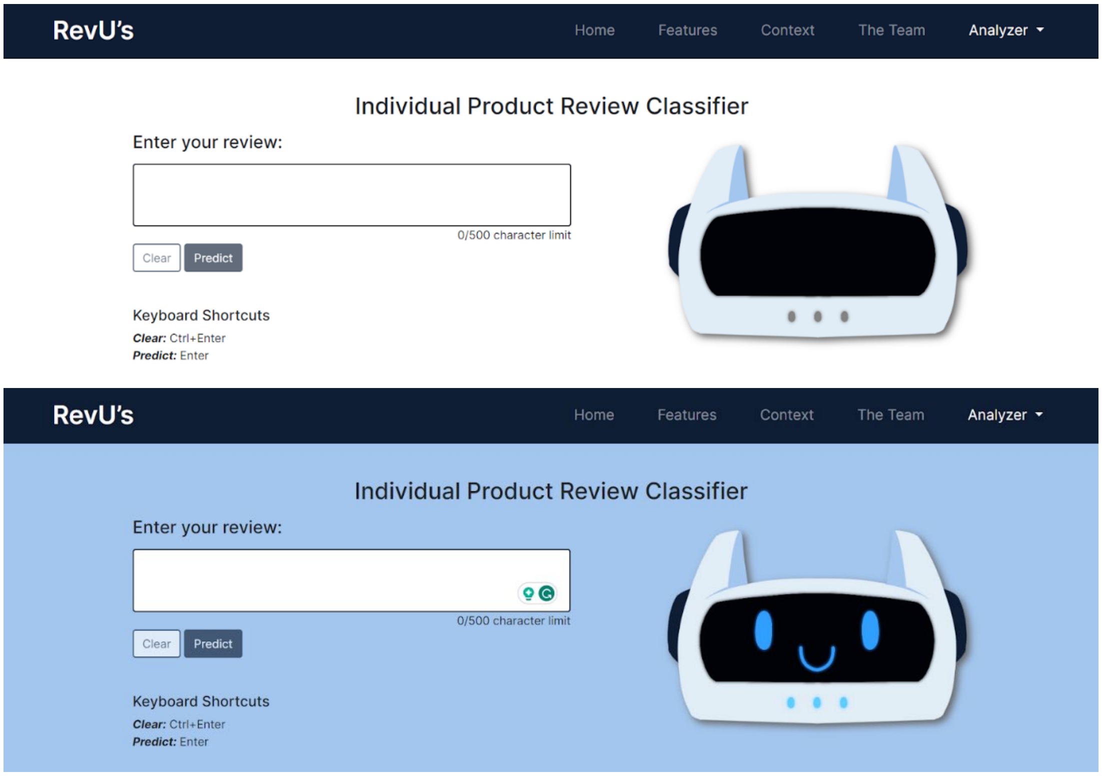

Figure 17. Individual Product Review Classifier Page

Tailored for individual product reviews, this page imposes a character limit, disabled buttons when the textbox is empty, and integrates keyboard shortcuts for a seamless user experience.

Figure 18. Consolidated Product Review Analyzer Page

Optimized for bulk reviews, this page introduces drag-and-drop functionality, disabled buttons when no file is selected, and dynamic visuals that respond to user actions such as mouse hover, file uploading, and the analysis process. This empowers users to efficiently analyze and draw insights from a consolidated set of product reviews.

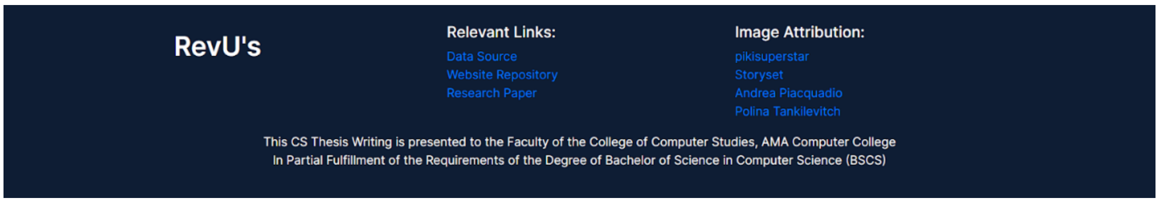

Figure 19. Footer Part for Disclaimer

Every page features a footer with a concise disclaimer, ensuring users are aware of the model's strengths and limitations. This adds transparency and sets expectations for the accuracy and applicability of the sentiment analysis.

Our application combines a visually appealing interface with intuitive functionalities, creating a seamless user sentiment analysis experience.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

Conclusion

The culmination of our efforts is reflected in the developed web application system, a powerful tool designed to support small businesses in decision-making by adeptly analyzing customer reviews and extracting invaluable insights for the enhancement of their offerings, ultimately benefiting their customers.

Foremost, our study has achieved the creation of a robust machine learning model proficient in classifying customer satisfaction into three fundamental categories: positive, negative, and neutral. This accomplishment, central to our study's objectives, fortifies the system's capacity to discern sentiments expressed across a diverse spectrum of product reviews. This functionality is seamlessly integrated into both the Individual Product Review Classifier and the Consolidated Product Reviews Analyzer.

Additionally, the objective to streamline the review process has been realized through the implementation of a user-friendly feature. The system empowers users by offering various methods for uploading files, allowing them to choose between the options of dragging and dropping files or selecting them through the traditional file selection method. This versatile approach not only enhances the efficiency of the input process but also reflects our commitment to providing a user-friendly experience, ensuring accessibility and ease of use. The system empowers users to provide a list of reviews, thereby enhancing the efficiency of the input process. This streamlined approach ensures a seamless and effective user experience, aligning perfectly with our study's goal to simplify the often intricate task of reviewing multiple product evaluations. Notably, this feature is particularly prominent in our Consolidated Product Reviews Analyzer, where users can input numerous reviews through the use of CSV files, receiving a comprehensive overall result.

Lastly, our study has successfully created functionality that delivers an overarching analysis of customer satisfaction based on a set of product reviews. The system transcends simple sentiment classification by extracting keywords or phrases that significantly contribute to sentiment analysis. This achievement mirrors our objective to provide comprehensive insights that extend beyond basic sentiment categorization.

In conclusion, each objective has been meticulously followed, answered, and implemented in our developed web application system, marking the system as an attestation to our commitment to advancing decision support for small businesses through the effective utilization of sentiment analysis on customer reviews.

Recommendations

To improve the web application's capabilities, we recommend scaling up its review processing capacity. This enhancement aims to equip the system with the ability to efficiently analyze a larger volume of reviews, facilitating more extensive datasets and ensuring a comprehensive analysis of customer sentiments.

For advancing the sophistication of sentiment analysis, exploration of advanced machine learning models is imperative. Specifically, addressing challenges related to mixed sentiments within a single review and incorporating multilingual capabilities are key considerations. Models adept at deciphering reviews with diverse sentiments and languages would significantly enhance the system's analytical prowess.

Diversifying the datasets used for training is crucial for improving the model's generalization capabilities. By incorporating datasets from a broader array of industries, regions, and demographics, we ensure the model's robust performance across varied product categories and user groups.

Continuous model training is recommended to keep the sentiment analysis model alongside of evolving language trends and user sentiments. Regular updates to the model ensure its adaptability to shifting linguistic patterns and changing expressions over time.

A critical aspect is evaluating and mitigating biases in the sentiment analysis model. A thorough assessment is essential, particularly regarding diverse datasets and user demographics. Implementing strategies to address biases ensures fair and accurate sentiment classification across different user groups.

Collaboration with linguistic and cultural experts can provide valuable insights into language-specific nuances and cultural variations. Integrating this expertise into the model training process enhances its understanding of diverse linguistic and cultural expressions.

Establishing a feedback loop with users is recommended for continuous improvement. Regular iterations based on user feedback will align the system's features with user expectations and requirements, ensuring a user-centric design.

Considering expanding the limit on the number of reviews processed will enhance the scalability of the system, accommodating more comprehensive analyses with larger datasets.

Integration of advanced grammar-checking algorithms is suggested for future developments. This addition would contribute to a more refined language analysis, enabling a nuanced understanding of customer sentiments within the reviews.

Improve user convenience in the Consolidated Product Reviews Analyzer by introducing a column selector feature for CSV file uploads. This enhancement allows users to specify the column containing reviews, offering flexibility beyond a fixed "Reviews" column title. This adjustment accommodates diverse data structures, ensuring seamless functionality and enhancing the application's adaptability to varying datasets.

Exploring the possibility of connecting or integrating with other websites can enhance the system's versatility. Assessing potential collaborations with external systems or services could enrich overall functionality without compromising the system's efficiency.

Collectively, these recommendations aim to bolster the effectiveness, versatility, and user experience of the sentiment analysis system developed in the thesis study.

REFERENCES

Alpaydin, E. (2020). Introduction to Machine Learning, Fourth Edition. United States: MIT Press.

Aggarwal, C. C., & Zhai, C. (2012). Mining Text Data. Springer Science & Business Media. doi.org/10.1007/978-1-4614-3223-4

Arthamevia, N. P., Adiwijaya, and Purbolaksono, M. D. (2021). Aspect-Based Sentiment Analysis in Beauty Product Reviews Using TF-IDF and SVM Algorithm. 2021 9th International Conference on Information and Communication Technology (ICoICT), pp. 197-201. doi:10.1109/ICoICT52021.2021.9527489

Beklemysheva, A. (2023). Why Use Python for AI and Machine Learning?. Steel Kiwi. Retrieved from https://steelkiwi.com/blog/python-for-ai-and-machine-learning/

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. Retrieved from https://link.springer.com/book/9780387310732

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357).

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine, 9(2), 48-57.* Retrieved from https://ieeexplore.ieee.org/document/6786458

Chan, N. (2023). Filipino Online Shopping Behavior: Statistics & Trends You Need to Know. Spiralytics. Retrieved from https://www.spiralytics.com/blog/filipino-online-shopping-behavior-statistics-trends/

Corpuz, R. S. (2021). Categorizing Natural Language-Based Customer Satisfaction: An Implementation Method Using Support Vector Machine and Long Short-Term Memory Neural Network. International Journal of Integrated Engineering. Retrieved from https://penerbit.uthm.edu.my/ojs/index.php/ijie

Crismundo, K. (2022). Most Pinoy e-shoppers read reviews first: report. Philippine News Agency. Retrieved from: https://www.pna.gov.ph/articles/1173283

Delos Santos, J. (2023). 51% of Pinoy businesses have a website, online store, and e-commerce accounts — study. NOYPIGEEKS. Retrieved from https://www.noypigeeks.com/business/pinoy-website-online-store-study/

Diekson, Z. A., Prakoso, M. R. B., Putra, M. S. Q., Syaputra, M. S. A. F., Achmad, S., & Sutoyo, R. (2023). Sentiment analysis for customer review: Case study of Traveloka. Procedia Computer Science, Volume 216, p. 682-690. doi:10.1016/j.procs.2022.12.184

Gorro, K., Feliscuzo, L., & Romana, C. L. S. (2022). Web Page Blocker using Machine Learning Classifier and Chrome Plugin. 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), pp. 1-5, doi: 10.1109/ICITDA55840.2022.9971159.

Goyenechea, A., Nartea, M., & Santos, J. J. (2019). Exploring the Characteristics of Millennials in Online Buying Behavior. Asia Pacific Journal of Multidisciplinary

Research, Vol. 7, No. 4, Part III. Retrieved from https://www.apjmr.com/wp-content/uploads/2020/01/APJMR-2019-7.4.03.01.pdf

Hossain, M. J., Joy, D. D., Das, S., & Mustafa, R. (2022). Sentiment Analysis on Reviews of E-commerce Sites Using Machine Learning Algorithms. 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), pp. 522-527. doi:10.1109/ICISET54810.2022.9775846

Ignacio, D. V. G. S., Sicangco, H. Y., & Esquive, J. A. (2023). Pop Music Midi Generation using Long Short-Term Memory based Recurrent Neural Network. Retrieved from https://ieomsociety.org/proceedings/2023manila/452.pdf

Irawaty, I., Andreswari, R., & Pramesti, D. (2020). Vectorizer Comparison for Sentiment Analysis on Social Media Youtube: A Case Study. 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), pp. 69-74. doi:10.1109/IC2IE50715.2020.9274650

Koubbi, R. (2022). A 2020 Guide to Python 2 vs Python 3. Retrieved from https://careerkarma.com/blog/python-2-vs-python-3/

Laato, S., Mäntymäki, M., Minkkinen, M., Birkstedt, T., Islam, A.K.M.N., & Dennehy, D. (2022). Integrating Machine Learning With Software Development Lifecycles: Insights From Experts. April 2022 Conference: European Conference on Information Systems. Retrieved from https://www.researchgate.net/publication/360318448_Integrating_Machine_Learning_With_Software_Development_Lifecycles_Insights_From_Experts

Machova, K., Mach, M., & Vasilko, M. (2021). Comparison of Machine Learning and Sentiment Analysis in Detection of Suspicious Online Reviewers on Different Type of Data. Sensors 2022; 22(1):155. doi:10.3390/s22010155

Mhatre, S. (2022). Sentiment Analysis with Python. Red Gate. Retrieved from https://www.red-gate.com/simple-talk/development/data-science-development/sentiment-analysis-python/

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill. Retrieved from https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf

Morgan Stanley. (2022). Here's Why E-Commerce Growth Can Stay Stronger for Longer. Retrieved from https://www.morganstanley.com/ideas/global-ecommerce-growth-forecast-2022

Nguyen, Heidi; Veluchamy, Aravind; Diop, Mamadou; and Iqbal, Rashed (2018) Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches. SMU Data Science Review: Vol. 1: No. 4, Article 7.Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss4/7

Ordinario, C. (2021). PHL online shoppers rely on ratings, reviews. BusinessMirror. Retrieved from https://businessmirror.com.ph/2021/10/12/phl-online-shoppers-rely-on-ratings-reviews/

Patel, S. (2023). A Guide to Customer Sentiment Analysis (and Why It Matters). REVE Chat. Retrieved from https://www.revechat.com/blog/customer-sentiment-analysis/

Perara, A. (2023). Leveraging E-commerce Customer Reviews for Increased Sales and Customer Loyalty. BusinessTechWeekly.com. Retrieved from https://www.businesstechweekly.com/digital-and-online-marketing/customer-experience/e-commerce-customer-reviews/

Prasanna, N. M., Mohan, S. R., Reddy, K. V. V., Kumar, B. S., Babu, C. G., & Priya, P. (2023). Machine Learning-Based Automated Disaster Message Classification System Using Linear SVC Algorithm. In J. Hemanth, D. Pelusi, & J. IZ. Chen (Eds.), Intelligent Cyber Physical Systems and Internet of Things. ICoICI 2022. Engineering Cyber-Physical Systems and Critical Infrastructures, 3. Springer. doi:10.1007/978-3-031-18497-0_63

Prastyo, P.H., Ardiyanto, I., & Hidayat, R. (2020). Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF. 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), pp. 1-6. doi:10.1109/ICDABI51230.2020.9325685

Prema, A. M., Hema, M. S., Maheshprabhu, R., & Nageswara, M. (2021). Sentimental Analysis of Twitter Data using Machine Learning Algorithms. 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), pp. 1-5, doi: 10.1109/FABS52071.2021.9702681

Reyes, R. R. (2021). ‘Sentiment analysis’ on Facebook reveals Filipinos’ online concerns amid pandemic. BusinessMirror. Retrieved from https://businessmirror.com.ph/2021/02/28/sentiment-analysis-on-facebook-reveals-filipinos-online-concerns-amid-pandemic/

Rokach, L., & Maimon, O. (2015). Clustering methods. Data mining and knowledge discovery handbook. Springer US. Retrieved from https://www.researchgate.net/publication/236005657_Data_Mining_and_Knowledge_Discovery_Handbook_2nd_ed

Sagarino, V. M. C., Montejo, J. I. M., & Ceniza-Canillo, A. M. (2022). Sentiment Analysis of Product Reviews as Customer Recommendations in Shopee Philippines Using Hybrid Approach. 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), pp. 1-6, doi: 10.1109/ICITDA55840.2022.9971379

Smith, A., & Jones, B. (2018). Advancements in Sentiment Analysis: Assessing Models with Cross-Platform Data. Journal of Computational Linguistics, 22(4), 567-584.

St John, M. & Freeman, V. (2023). Python 2 Vs. 3: Choosing Which To Learn. Retrieved from https://www.forbes.com/advisor/education/python-2-vs-python-3/

Vijayaragavan, P., Ponnusamy, R., & Aramudhan, M. (2020). An optimal support vector machine based classification model for sentimental analysis of online product reviews. Future Generation Computer Systems, Volume 111, p. 234-240. doi:10.1016/j.future.2020.04.046

Wolff, R. (2020). Sentiment Analysis & Machine Learning. MonkeyLearn. Retrieved from https://monkeylearn.com/blog/sentiment-analysis-machine-learning/

Xia, H., Yang, Y., Pan, X., Zhang, Z., & An, W. (2019). Sentiment analysis for online reviews using conditional random fields and support vector machines. Electron Commerce Ressearch 20, 343–360 (2020). doi:10.1007/s10660-019-09354-7

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530.

Zhang, H., Yu, D., Xu, T., & You, Y. (2019). A Survey on Challenges of Machine Learning Model Robustness. arXiv preprint arXiv:1908.02806.

APPENDICES

Appendix A. 1-Dataset_Prep.ipynb

Dataset Preparation

Dataset Loading

Load the CSV files into Data Frame

```
import os
import pandas as pd
file_path = os.path.join(os.getcwd(),'datasets/evaluators/')
df = pd.read_csv(file_path+'shopee_ProductReviews.csv') #Replace the dataset used
df.head()
df.shape
```

Field Selection

Select only the fields that will be used as the labeled data.

We only need the actual text (as the input feature) and rating (as the quantitative labels). If your dataset uses different column names for these attributes, please adjust the parameters accordingly.

```
df = df.rename({"text" : "Reviews"}, axis=1)
df = df.rename({"label" : "Score"}, axis=1)
df = df [["Score", "Reviews"]]
df.head()
```

Missing and Duplicate Data Checks

Eliminate any missing or duplicate data that could introduce bias and impact the model's true accuracy.

```
def missing_checker(df):
    missing_reviews = df.isnull()

    print("Missing data:\n", missing_reviews.sum())
    print(df[missing_reviews.any(axis=1)])
missing_checker(df)
```

```
# Drop Missing Data
df.dropna(inplace=True)

# Check again
missing_checker(df)
```

```
def duplicate_checker(df):
    # Get the duplicate reviews
    duplicate_reviews = df.duplicated(subset='Reviews')

    print("Duplicate Reviews:\n", duplicate_reviews.sum())

    duplicate_rows = df[duplicate_reviews]
    print(duplicate_rows['Reviews'])
duplicate_checker(df)
```

```
# Remove the duplicates but keep the first instance
df.drop_duplicates(subset='Reviews', keep='first', inplace=True)
duplicate_checker(df)
```

Feedback Mapping of Scores

In this step, we are transforming numerical ratings into qualitative labels - 'Positive,' 'Neutral,' and 'Negative.' These labels represent the sentiment associated with the input data, and this transformation facilitates sentiment analysis and classification. The resulting 'Feedback' column will contain the mapped sentiment labels for each data point which is the actual label.

```
# Ensure that Score only contains int
df['Score']  = pd.to_numeric(df['Score'], errors='coerce')
df['Score']  = df['Score'] .fillna(0).astype(int)
```

```
import numpy as np                       # MD array and Matrices

conditions = [
    (df['Score'] >= 4),
    (df['Score'] == 3),
    (df['Score'] <= 2)
    ]

feedback_values = ['Positive','Neutral',’Negative']
df['Feedback'] = np.select(conditions, feedback_values)

feedback_counts = df['Feedback'].value_counts()
print(feedback_counts)
```

```
import matplotlib.pyplot as plt  # Data Visualization

df['Feedback'].value_counts().sort_index().plot.bar(color=['maroon', 'steelblue',
'limegreen'])
```

```
print("Total number of reviews:",df.shape[0])
df.head()
```

Data Preprocessing

Clean the input text so that the algorithm can handle the data better. The actual code for this is in code/data_preprocess.py

```
import sys
sys.path.append('../code/')
from data_preprocess import text_cleaner

df['Reviews'] = df['Reviews'].apply(text_cleaner)
```

```
# Check for missing and dupe data again isn't substantial that text cleaner wiped it
out.
missing_checker(df)
duplicate_checker(df)
```

```
# Drop
df.dropna(inplace=True)
missing_checker(df)
df.drop_duplicates(subset='Reviews', keep='first', inplace=True)
duplicate_checker(df)
```

```
df.head()
df.shape
```

```
# Save the DataFrame to the new CSV file
df.to_csv(os.path.join(file_path, 'prep_shopee_ProductReviews.csv'), index=False,
encoding='utf-8')
```

PROCEED TO MODEL TRAINING

Appendix B. data_preprocess.py

```
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords', quiet=True)


stop_words = set(stopwords.words('english'))
negative_words = {'no', 'not', "never", "neither", "nor", "none", "nobody",
"nowhere", "nothing", "hardly", "scarcely"}
stop_words -= negative_words # Remove the Negative words


import contractions
import re

def text_cleaner(text):
    text = str(text).lower()                   # Convert to string and lowercase
    text = contractions.fix(text)              # Expand Contractions
    text = re.sub(r'<.*?>', ' ', text)         # Remove HTML tags
    text = re.sub(r'http\S+', ' ', text)       # Remove URLs using regular expression
not covered by tags
    text = re.sub(r'[^A-Za-z0-9\s]', ' ', text) # Remove non-alphanumeric characters
using regular expression
    words = text.split()                       # Tokenize for stop word removal
    cleaned_text = " ".join(word for word in words if word not in stop_words) #
Remove the stopwords from the text.
    print(cleaned_text)
    return cleaned_text
```

Appendix C. 2-SVM_ModelTrain.ipynb

Dataset Loading

```
import os
import pandas as pd

review_dir = os.path.join(os.getcwd(), 'datasets/prep_reviews.csv')
load_df = pd.read_csv(review_dir)
load_df.dropna(inplace=True)     #Drops null data that csv added during save

print(load_df.shape)
load_df.head()
```

```
feedback_counts = load_df['Feedback'].value_counts()
print(feedback_counts)
```

Resampling Technique

Implement resampling techniques to downsize the training dataset when larger models do not significantly enhance accuracy or fail to maintain a reasonable file size.

This is crucial to ensure that the model's performance gains justify the increase in file size, as excessively large files may result in slower loading times.

```
my_sample=.20
df = load_df.sample(frac=my_sample, random_state=42)
print(f"Dataset       : {load_df.shape[0]}")
print(f"Sampled       : {df.shape[0]}({my_sample*100}% of Dataset)\n")
```

Dataset Splitting

The dataset is split into two separate sets:

The training set is used to train the machine learning model

The test set is used to evaluate its performance on unseen data.

```
# Map the Data
x = df['Reviews']  # Features (processed text)
y = df['Feedback']  # Target labels (Positive-Neutral-Negative)
```

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,
random_state=42)

print("Sampled Reviews:")
print(f"Training set  : {len(x_train)}")
print(f"Testing set   : {len(x_test)}")
print(f"Total         : {df.shape[0]}")
```

Vectorization

Convert the text data into numbers so that the SVM algorithm can understand.

This conversion is like turning words into numbers, making it easier for the algorithm to analyze text and provide accurate feedback classification and customer satisfaction analysis."

```
%%time
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(ngram_range=(1, 2))
vectorizer.fit(x_train)

x_train_vectorized = vectorizer.transform(x_train)
x_test_vectorized = vectorizer.transform(x_test)
```

Model Training

We use sklearn library's SVM algorithm to train a model using the training data.

The SVM model learns from the labeled data to classify reviews into their respective sentiment categories (positive, neutral, or negative).

```
%%time
from sklearn.svm import LinearSVC
classifier = LinearSVC()

# Train the Model
classifier.fit(x_train_vectorized, y_train)

# Make predictions on the test set
y_pred = classifier.predict(x_test_vectorized)
```

Save the Model & Evaluation

Save and evaluate the model using the metrics.

Feel free to adjust the code/methods to your liking to create a better model.

```
import joblib
```

```
save_path = os.path.join(os.getcwd(), 'models/test_model/')
classifier_path = save_path+"SVM_classifier.joblib"
vectorizer_path = save_path+"tfidf_vectorizer.joblib"

#Save the Model
joblib.dump(classifier, classifier_path)
joblib.dump(vectorizer, vectorizer_path)
```

```
# Get the file size
vectorizer_size = os.path.getsize(classifier_path) / 1024  # Convert to MB
classifier_size = os.path.getsize(vectorizer_path) / 1024  # Convert to MB
```

```
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Evaluate the metrics
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)

# Print the metrics to the console
print("Accuracy:", accuracy)
print("Confusion Matrix:\n", conf_matrix)
print("Classification Report:\n", classification_rep)

# Save the metrics
eval_path = os.path.join(os.getcwd(), 'model_evaluations/0_evaluation_results.txt')
with open(eval_path, 'w') as file:
    file.write("Model Specifications\n")
    file.write(f"Classifier    : {classifier} \n")
    file.write(f"Vectorizer    : {vectorizer} \n")
    file.write(f"Dataset       : {load_df.shape[0]}\n")
    file.write(f"Sampled       : {df.shape[0]}({my_sample*100}% of Dataset)\n")
    file.write(f"Training set  : {len(x_train)} \n")
    file.write(f"Testing set   : {len(x_test)} \n\n")
    file.write(f"File Sizes\n")
    file.write(f"Vectorizer   : {vectorizer_size:.2f}KB \n")
    file.write(f"Model        : {classifier_size:.2f}KB \n\n")


    file.write("Model Evaluation\n")
    file.write(f"Accuracy      : {accuracy*100} \n\n")
    file.write(f"Confusion Matrix:\n{conf_matrix} \n\n")
    file.write(f"Classification Report:\n{classification_rep} \n")

print("Evaluation results saved to:", eval_path)
```