Procurement Control Tower:
Proof of Concept through Machine Learning and Natural Language Processing

by

Bishwajit Kumar
Bachelor of Mechanical Engineering, Sathyabama Institute of Science & Technology, India, 2007

and

Pablo Andrés Barros Gómez
Master of Business Administration, Nyenrode Business University, The Netherlands, 2019

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2023

Signature of Author: ____
Department of Supply Chain Management
May 12, 2023

Signature of Author: ____
Department of Supply Chain Management
May 12, 2023

Certified by: ____
Dr. Elenna Dugundji
Research Scientist, MIT Center for Transportation and Logistics
Capstone Advisor

Certified by: ____
Dr. Thomas Koch
Postdoctoral Associate, MIT Center for Transportation and Logistics
Capstone Co-Advisor

Accepted by: ____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Procurement Control Tower:
Proof of Concept through Machine Learning and Natural Language Processing

by

Bishwajit Kumar
Bachelor of Mechanical Engineering, Sathyabama Institute of Science & Technology, India, 2007

and

Pablo Andrés Barros Gómez
Master of Business Administration, Nyenrode Business University, The Netherlands, 2019

Submitted to the Program in Supply Chain Management
on May 12, 2023, in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

An organization's procurement process is pivotal for its success in a competitive market. The increased uncertainty and complexity of post-pandemic supply chains have made procurement a more valuable focus point among organizations and a differentiating factor to achieve a competitive advantage. The sponsor company of this study believes that the key to being competitive in today's VUCA (volatile, uncertain, complex, and ambiguous) market relies on getting faster insights into the problem areas, having enhanced decision-making capabilities, and optimal exception management. For that reason, it seeks to understand if said competencies are encapsulated within a Procurement Control Tower's value proposition. To meet our sponsor's requirements, we divided our research into two components, a qualitative and a quantitative component. The first evaluates and defines the scope, value proposition, and deployment strategy of the Procurement Control Tower. The latter provides proof of concept by creating a working prototype of one of its use cases. The selected use case for the prototype is Spend Analytics, more specifically, the categorization and sub-categorization of the unclassified spend data for an assigned business unit. To create the prototype, this study compares multiple Machine Learning algorithms and selects Random Forest as the best-performing one in terms of accuracy. The algorithm's predictive power is then enhanced by pre-processing the data with Natural Language Processing. The final model performs with 94% accuracy at a category level and 90% at a sub-category level. This study's primary finding, obtained through the categorization of approximately 250 million USD of unclassified spend data, is that implementing the Procurement Control Tower in the sponsor's business provides measurable value. For our assigned business unit, it creates renegotiation opportunities with suppliers, increases budgeting accuracy, and reduces the man-hours required. The final algorithm of the prototype has been presented to the sponsor company, which is currently deploying it for the assigned business unit. To scale up the benefits of the solution across the organization, the sponsor plans to deploy it for the remaining business units.

Capstone Advisor: Dr. Elenna Dugundji
Title: Research Scientist, MIT Center for Transportation and Logistics
Capstone Co-Advisor: Dr. Thomas Koch
Title: Postdoctoral Associate, MIT Center for Transportation and Logistics

ACKNOWLEDGMENTS

We are thankful to our advisor Dr. Elenna R. Dugundji for mentoring and guiding us throughout this capstone. Thank you for your commitment and effort put into this research. We would also like to thank our co-advisor Dr. Thomas Koch for the technical guidance. A special thanks to Toby Gooley for her constant motivation and inspiration during the whole process. We would like to thank our sponsors for trusting us and being regularly available for guidance. To all the SCM staff and colleagues at the MIT CTL, thank you for the great support you have provided throughout the program.

Bishwajit

I would like to thank my parents, Sneh Lata and Suresh Prasad, for their sacrifices to support me through every phase of my life. Special thanks to my wife, Shubhrika, for motivating me throughout the MicroMaster's journey and my time at MIT. You are the best! I am thankful to my cute daughters, Saanvi, and Saavi, for the unconditional love you offered! To my wonderful Capstone partner, Pablo Barros - working with you was a fun experience. I've enjoyed working together and chatting about interesting topics for long hours.

Pablo

I wish to extend boundless gratitude to those without whom this Capstone would not have come to fruition. The unconditional support, and unwavering encouragement of Doris Gomez, Jose and Laura Barros, Walter Bayona, and the SD association. Their life-long upholding and acumen are the foundation and inspiration of this research. Special acknowledgment to my Capstone colleague Bishwajit, our year-long partnership has been a highly rewarding one.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1 INTRODUCTION

1.1 Motivation

An organization's procurement process is pivotal for its success in a competitive market. As companies scale up in magnitude, they allocate more resources and effort to leverage the opportunities it offers by optimizing the way they obtain goods and services from their suppliers. The increased uncertainty and complexity of post-pandemic supply chains have made procurement a more valuable focus point among organizations and a differentiating factor to achieve a competitive advantage.

Such is the case of our capstone sponsor company whose large annual spending of over 30 billion USD calls for an in-depth analysis of its procurement strategies to efficiently allocate this budget. Out of this annual spending, more than 10 billion USD is designated for the sourcing of materials through the supply chain, with the rest being attributed to ends such as information technology, marketing, and R&D.

The supply chain of our sponsor is complex. It mainly works in three distinct sectors. Owning dozens of brands among these three sectors makes the types of products that our sponsor sources very varied in nature. It has an extensive supplier base that surpasses 45,000 individual suppliers globally. It uses several different supporting technologies to manage its supply chain data. Some examples of these supporting technologies are SAP, E2E visibility platforms, and different versions of Azure Tools. The proliferation of supporting technologies has led to fragmented communication among them, due to the lack of a comprehensive system that can unify the insights derived from multiple sources. All in all, our sponsor's large spending, highly varied product needs (ranging from chemicals to rare metals and semiconductors), extensive supplier base, diverging software insights, and overall presence in multiple business sectors create a deeply layered procurement system.

Our sponsor believes that the key to being competitive in today's VUCA (volatile, uncertain, complex, and ambiguous) market relies on getting faster insights into the problem areas, having enhanced decision-making capabilities, and optimal exception management. For that reason, it seeks to understand if said competencies are encapsulated within the Procurement Control Tower's value proposition.

1.2 Problem Statement and Research Questions

Control towers are digital platforms that “capture and use near real-time operational data from across the business ecosystem to provide enhanced visibility and improve decision-making.” (Gupta, 2022) Our sponsor has already implemented two control towers in areas other than procurement, namely the planning and the delivery areas. At the same time, it deems the use of cutting-edge technologies paramount to enhance its procurement strategies. The company’s strong capabilities for implementing control towers coupled with, in its view, a sub-optimal system to guide its procurement decision-making creates an unfulfilled opportunity that it seeks to capitalize on. This study assists our sponsor in formulating a roadmap to pursue said opportunity.

Research Questions

I. Will a Procurement Control Tower create measurable value for the sponsor’s procurement functions?

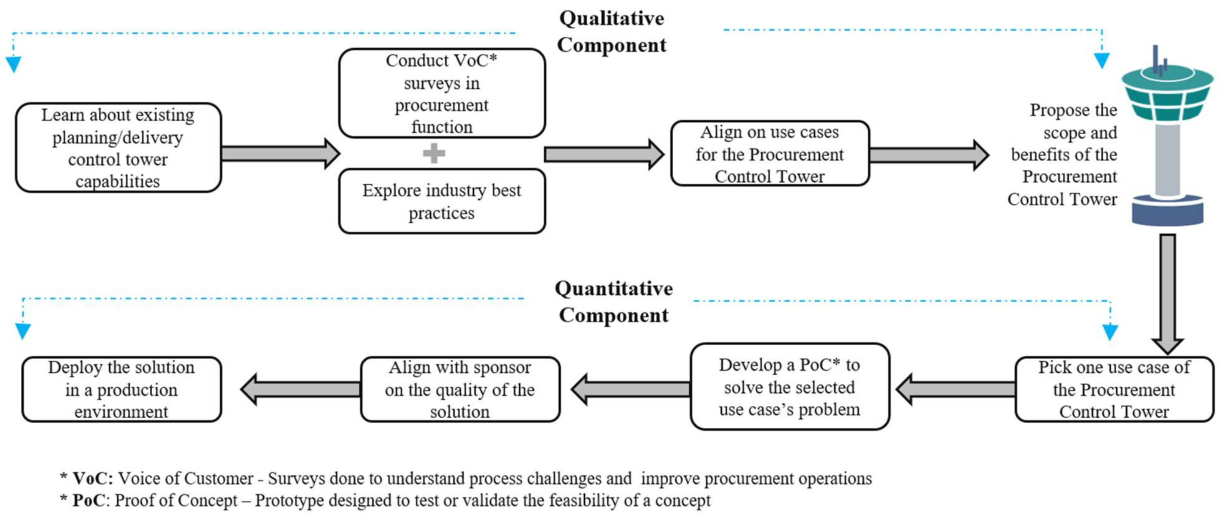
II. Can we demonstrate proof of value of a Procurement Control Tower by creating a prototype of one of its use cases?

1.3 Scope Definition

Our plan of work is explained in detail in Figure 1. To meet our sponsor’s requirements, we divided our research into two components, a qualitative component that evaluates the overall value proposition of a Procurement Control Tower, and a quantitative component that provides proof of concept demonstrating measurable value of the Procurement Control Tower.

Figure 1

Flow Chart of Plan of Work



Qualitative Component
Learn about existing planning/delivery control tower capabilities
Conduct VoC* surveys in procurement function
Explore industry best practices
Align on use cases for the Procurement Control Tower
Propose the scope and benefits of the Procurement Control Tower
Quantitative Component
Pick one use case of the Procurement Control Tower
Develop a PoC* to solve the selected use case's problem
Align with sponsor on the quality of the solution
Deploy the solution in a production environment

* **VoC:** Voice of Customer - Surveys done to understand process challenges and improve procurement operations
* **PoC**: Proof of Concept – Prototype designed to test or validate the feasibility of a concept

Our qualitative research began by conducting interviews with the subject matter experts of the planning and delivery control towers. We studied the control towers' capabilities and limitations. Then we discussed the pain points of the various procurement functions with the relevant subject matter experts through VoC (Voice of Customer) surveys. We also compared this feedback with what's being done in the industry as best practices. We then proposed the benefits that a control tower would bring to the procurement area through a series of use cases.

Our quantitative research began by selecting one of the proposed use cases of the Procurement Control Tower to focus on. After narrowing the scope, we developed a prototype, in a sandbox environment, of this use case as a PoC (Proof of Concept). The prototype offers a solution that shows measurable value and can be the starting point and foundation of a full-scale Procurement Control Tower.

To select the use case to be used for prototyping, we established a short list of six key use cases in which the Procurement Control Tower would bring the most value to the area. The following is the list of use cases, together with a short description of the value that a control tower would bring to them:

1. Contract Management: The automation and streamlining of the processes that the contract management team does regularly would optimize the time resource of the personnel assigned to these tasks. It would also reduce the turnaround time of contract processing and the chances of human error.
2. Supplier-Enabled Innovation: The control tower could have real-time data about the suppliers in a way that allows for evaluating the degree of innovation they achieve. This information could be used then by the sponsor to focus their attention and/or investment on suppliers that they deem valuable in terms of innovative potential.
3. New Supplier Performance Management: Our sponsor is regularly working with new suppliers, and the performance of these could be measured more optimally by leveraging the power of advanced data analytics and/or artificial intelligence.
4. Supplier Risk Management: Currently, our sponsor is periodically revising the risk indicators of their suppliers in a semi-automated way. The digital control tower could have the capability to process the data of the current suppliers in real-time and provide insights for decision-making. This would help prevent supplier-related issues by noticing patterns that indicate a future complication before these patterns can be noticed by the human eye.
5. Sourcing: The digital control tower could help the sourcing of materials by providing real-time visibility of these, by using predictive analytics to prevent disruptions that may impact the sourcing of materials, and by allowing members of different areas of the organization to converge in one single platform that shows every member the same state of the data.
6. Spend Analytics – Spend Categorization: The database that the procurement team uses to analyze the company's purchased materials has a sizable number of records that are not labeled to indicate the category or sub-category of materials which they belong to. The control tower has the potential to automatically categorize this unclassified spend data and provide a deeper level of insights from their data analysis.

Together with the sponsor company, we decided to select **Spend Analytics - Spend Categorization** (of materials) as the use case to show measurable value for the Procurement Control Tower. The main reason for this selection is that a significant number of the spend records lack accurate labels indicating the category or sub-category of materials they belong to. This makes any report or analysis done based on categories of spending data increasingly inaccurate. The data provided to us would be the 2022 spend data records of one assigned business unit of the company which has approximately **250M USD** of unclassified spend data annually.

1.4 Hypothesis

We hypothesized that, after thoroughly understanding our sponsor's pain points and selecting the categorization of unclassified spend data as the use case to show measurable value for the Procurement Control Tower, we would develop a working prototype that categorizes real business data through the use of machine learning, a subset of artificial intelligence. To develop this prototype, we would compare an array of different algorithms to select the best-performing one and thus, the most suitable to accomplish the goal. This prototype would then exemplify the value of a full-scale Procurement Control Tower.

1.5 Objectives and Expected Deliverables

The objective of the project is to provide a foundation and/or a baseline for the sponsor to follow in its effort to optimize, innovate, and streamline its procurement processes. We must demonstrate the inherent, measurable value that a Procurement Control Tower will provide as the de facto tool for the procurement team to manage its source-to-pay processes. To do so, there is a set of expected deliverables that must be presented.

<u>Deliverables</u>:

I. An architecture that defines what a digital control tower of the sponsor's procurement function would mean and the value proposition it has.

II. A working prototype, made in a sandbox environment, of a valuable use case of the Procurement Control Tower. This prototype would use near real-time data to show the benefits that implementing a control tower would bring for the procurement area. It should also, in turn, act as a foundation for the sponsor company to expand and develop a full-scale version that enhances the insights for procurement decision-making and exception management.

1.6 Conclusion

This research paper is divided into two components, a qualitative component, and a quantitative component. The first engages in defining what a digital control tower would mean for the procurement area of our sponsor company by designing an architecture that reflects its value proposition. The latter engages in showing measurable value for the Procurement Control Tower by developing a prototype of one of its use cases/capabilities. The selected use case is the enhancement of the Spend Analytics. To develop this prototype, this study considers the applications of machine learning, a subset of artificial intelligence,

The results of these two components support the decision of whether the sponsor company should invest in the development and implementation of a full-scale control tower for the procurement area, as it already has in other areas of the organization. This could ultimately aid our sponsor's procurement team in their daily functions to make timely decisions leading to improved profitability, and better stakeholder satisfaction.

2 STATE OF THE ART

Our capstone is divided into a qualitative component and a quantitative component. To address the qualitative component, we review the literature on the following subjects:

2.1. Defining Supply Chain Control Towers

2.2. Key Characteristics of Control Towers

2.3. Software Implementation Strategies.

To address the quantitative component, we review the literature on the following subjects:

2.4. Comparative Matrices of Categorization Machine Learning Literature

2.5. Machine Learning Algorithms for Categorization

2.6. Pre-Processing Techniques

2.1 Defining Supply Chain Control Towers

Control Towers have been defined in many different ways by different sources, to have a clear grasp of their meaning we look at various definitions and conclude our own.

According to Gupta (2022), the definition of a control tower is a complex topic in the market, but he states that control towers combine different types of resources from all across an organization to enhance the visibility of its processes and consequently its decision-making.

Christian Titze, VP Analyst at Gartner says that control towers combine information from various applications or tools across a company into a single place to come up with insights, predictions, and suggestions that wouldn't have been accessible otherwise. Gupta (2022)

SAP (2022) says that control towers are cloud-based and use advanced technologies to produce insights to manage supply chains. It mentions the visibility benefits that control towers provide across multiple members of a company's supply chain and the fact that they're capable to help with cases such as preventing future issues and facilitating repetitive work. SAP also believes that past ways of handling supply chain issues, like demand forecasting and disruption management, are no longer appropriate, and that control towers offer a better and more future-proof way of dealing with these.

We conclude that control towers are digital tools that combine different types of information from various sources within an organization to simplify the understanding of how the data interacts with each other. This leads to enhanced insights for better decision-making. At the same time, we see that control towers are considered a modern and more efficient way to deal with supply chain issues.

Benefits of Supply Chain Control Towers

According to SAP (2022), the main benefits of control towers are:

- Providing end-to-end visibility - This simplifies the management of the supply chain by allowing to see real-time data about other parties in a company's supply chain.
- Providing tools for better decision-making - Control towers help translate complex and extensive data into layman's term so the insights are easier to share and understand.
- Making the supply chain more agile - The enhanced insights that control towers provides allow for a quicker and more efficient supply chain.
- Better supply chain collaboration - Since the data and insights are easier to share among both internal and external parties, teamwork sees a great benefit from control towers.
- Better inventory levels - Control towers help with better predictability across the supply chain, this allows inventory managers to enhance the accuracy of their forecasts and consequently reduce costs.

From these we conclude that when a control tower is successfully implemented it can generate an array of different benefits that can ultimately increase a company's bottom line as well as stakeholder satisfaction.

2.2 Key Characteristics of Control Towers

Control towers can greatly differ from company to company, and even from area to area. This is the case of our sponsor company, whose control tower of the planning area greatly differs from their control tower in the delivery area. Nonetheless, certain basic attributes allow for a control tower to do its assigned task. The following are the basic characteristics commonly shared by all digital control towers according to SAP (n.d.):

- Clean and complete data is the most important thing a control tower needs to generate insights that can enhance decision-making. The robustness and exhaustiveness of the data inputted is directly proportional to the quality of the results that the control tower could produce.
- Control towers join vastly different types of data into a single location to make sense of the relationship among them. SKU barcodes and weather data are examples of varying data types that could be joined in the control tower.
- Control towers rationalize the convoluted data they receive and create a simplified version of visualizing it. This allows high-level decision-makers to easily have real-time information about operational areas.
- Control towers have a predictive function, which anticipates risk and issues to allow teams to either prevent or prepare for them. Due to their granular visibility capabilities, they are also capable of quickly identifying and labeling problematic areas at an SKU level.
- Using advanced technologies like artificial intelligence or machine learning, control towers offer automated capabilities that greatly reduce manual labor in the organization and continuously self-learn to improve the quality of their output.
- Control towers make it easy for stakeholders to create procedure playbooks that define how to solve specific supply chain issues. These playbooks can be shared with supply chain teams to streamline and standardize the response when a new issue is detected.

When control towers become more sophisticated, they can involve a new level of more valuable characteristics. The following is a list of characteristics that sophisticated control towers commonly possess according to Verwijmeren (2017):

- **Multi-Party Supply Chain Support:** Firms have been increasingly incorporating more parties in their supply chains making them more layered and convoluted. As the number of parties increases in the same supply chain, the discrepancies and miscommunication among these become a bigger issue. This is where control towers can leverage their unifying capabilities and design a platform that will mitigate these risks.
- **End-to-End Visibility:** Visibility is always related to control towers, but end-to-end visibility takes it one step further to create insights at a more granular level. Beyond simple track and trace data, end-to-end visibility is capable of identifying key data about specific steps within a product's supply chain.
- **Real-Time Exception Management:** Control towers can continuously monitor business operations and processes in real-time. This allows the user to detect and address any exceptions as soon as they occur.
- **Order Orchestration and Optimization**: Control towers can leverage algorithms that forecast the optimal amount of product to order at any given time to reduce costs and inefficiencies.

2.3 Software Implementation Strategies.

Thomas H. Davenport (1998), in his Harvard Business Review article *Putting the Enterprise into the Enterprise System*, argues that the implementation of new systems in an organization requires careful deliberation and participation from top management. He also mentions the importance of defining a structured strategy that establishes the parameters of the implementation such as the number of functions that will be involved and the order in which the modules of the system will get implemented. This is because the implementation of a new system in an organization involves an inherent degree of risk. The risks could involve a multitude of different factors like data loss, security vulnerabilities, and implementation failure.

The cost to implement software could be many times the cost of the software itself. For this reason, having an adequate implementation strategy can greatly affect the ROI (Return on Investment) of the system. (Khanna & Arneja 2012).

According to Khanna & Arneja (2012), there are five main types of implementation strategies that we might use for new software in an organization. These are:

- **Big Bang Approach:** The Big Bang approach implements all software capabilities across all the business units of the organization at the same time. This strategy requires a detailed plan of action and a large number of resources invested for the implementation. Positive aspects of the Big Bang Approach are that the shift to the new system is done quickly and that the entirety of the organization is working at the same time to adapt it. Negative aspects of it are that it has a high risk of failure/complications and that it generally requires high investments of resources such as time and capital.
- **Phased Approach:** The phased approach is a multi-step strategy to implement a new system. The goal of this approach is to divide the system into sub-elements and implement them in sequential order. Each sub-element must be able to work independently for this approach to work. If sub-elements are intended to work in collaboration, the integration of these must be done at later stages of the process. The benefits of a phased approach include flexibility, cost reduction, and opportunity for feedback and correction.
- **Parallel Approach:** The parallel approach is based on implementing the new system, and keeping the older system (called the Legacy System) running simultaneously. The use of the Legacy System should be discontinued in due time but the timeframe on which the company will run both systems simultaneously is up to the user. Positive aspects of the parallel approach are that it is highly safe since it always has a backup system in case the new one faces difficulties and it allows for easy improvements to the new system based on feedback. The main negative aspect of this approach is that it requires a large number of resources to run both systems simultaneously. This approach is commonly used when the system is highly crucial and a malfunction of it would result in severe consequences to the operation.
- **Process Line Approach:** The process line approach consists in implementing the entirety of the system, but only on one specific process line of the organization. For example, implementing the

system first on the finance process line, then on the human resources process line, and then on the procurement process line. Under this approach, the less complex process lines are usually undertaken first to receive feedback and improve the implementation of the more complex process lines.

- **Hybrid Approach:** The hybrid approach is a combination of two or more of the previously explained approaches. This approach is more commonly used by larger corporations to satisfy and capture the complexity of their processes.

2.4 Comparative Matrices of Categorization Machine Learning Literature

In this research, we review the available literature to study the techniques and machine learning algorithms that have been previously utilized to achieve a similar objective to our own, which is to classify unseen data into pre-established categories and sub-categories. To do so we compare 20 different articles by the type of data that was classified and the methods utilized to classify it. The comparative analysis of these 20 articles will be illustrated in four matrices (Tables 1, 2, 3, and 4).

The first matrix (Table 1) states the reference of the 20 articles, the article title, and the topic of the data that the study classifies.

Table 1

Matrix of Literature Review: Data Topics

#	Reference	Article Title	Data Topic
1	Harikrishnakumar et. al (2019)	Supervised Machine Learning Approach for Effective Supplier Classification.	Suppliers
2	Chow (2022)	Categorizing Short Text Descriptions: Machine Learning or Not?	Spend
3	Jain (2019)	Comparative Study of Machine Learning Algorithms for Document Classification	Movie Reviews
4	Shah et al. (2020)	A Comparative Analysis of Logistic Regression, Random Forest, and KNN Models for the Text Classification	News Articles
5	Kambar et al. (2021)	Clinical Text Classification of Alzheimer's Drugs' Mechanism of Action	Clinical Text
6	Bangyal et al. (2021)	Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches	News Articles
7	Wang et al. (2019)	A clinical text classification paradigm using weak supervision and deep representation	Clinical Text
8	Padurariu & Breaban (2019)	Dealing with Data Imbalance in Text Classification	Work Experience
9	Arkok & Zeki (2020)	Classification of Holy Quran Verses Based on Imbalanced Learning	Religious Verses
10	Ali et al. (2019)	Adam Deep Learning with SOM for Human Sentiment Classification	Social Media Text
11	Occhipinti et al. (2022)	A pipeline and comparative study of 12 machine learning models for text classification	E-mail text
12	Qi (2020)	The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model	Theft Crime Data
13	Taha & Yousif (2023)	Enhancement of text categorization results via an ensemble learning technique	News Articles
14	Deng et al. (2018)	Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods	Clinical Text
15	Dubey et al. (2022)	Breast Cancer Modeling and Prediction Combining Machine Learning and Artificial Neural Network Approaches	Clinical Text

#	Reference	Article Title	Data Topic
16	Gaurnav et al. (2021)	CodeScan: A Supervised Machine Learning Approach to Open-Source Code Bot Detection	Code
17	Endalie & Haile (2021)	Automated Amharic News Categorization Using Deep Learning Models	News Articles
18	Parida et al. (2021)	News Text Categorization using Random Forest and Naïve Bayes	News Articles
19	Kamath et al. (2018)	Comparative Study Between Traditional Machine Learning and Deep Learning Approaches for Text Classification	Clinical Text
20	Suneera & Prakash (2020)	Performance Analysis of Machine Learning and Deep Learning Models for Text Classification	News Articles

The second matrix (Table 2) compares the pre-processing techniques performed on the data to enhance the predictions of the algorithms. The techniques compared in the matrix are Term Frequency – Inverse Document Frequency (**TF-IDF**), **Word2Vec**, **Doc2Vec**, Bag of Words (**BoW**), Bidirectional Encoder Representations from Transformers (**BERT**), and Tokenization (**TOKEN**). All of the pre-processing techniques are in the field of Natural Language Processing which will be further explained and discussed in Section 2.5. Note that not all studies use pre-processing techniques.

Table 2

Matrix of Literature Review: Pre-processing Natural Language Processing Techniques

#	Reference	Pre-processing: Natural Language Processing Techniques					
		TF-IDF	**Word2Vec**	**Doc2Vec**	**BoW**	**BERT**	**TOKEN**
1	Harikrishnakumar et. al (2019)						
2	Chow (2022)						
3	Jain (2019)						
4	Shah et al. (2020)	X					
5	Kambar et al. (2021)	X					
6	Bangyal et al. (2021)	X					X
7	Wang et al. (2019)	X		X			
8	Paduraru & Breaban (2019)	X			X		
9	Arkok & Zeki (2020)	X					
10	Ali et al. (2019)	X					X
11	Occhipinti et al. (2022)	X					
12	Qi (2020)	X					
13	Taha & Yousif (2023)	X					
14	Deng et al. (2018)						
15	Dubey et al. (2022)						
16	Gaurnav et al. (2021)						
17	Endalie & Haile (2021)						X
18	Parida et al. (2021)	X					X
19	Kamath et al. (2018)						
20	Suneera & Prakash (2020)	X	X			X	

From Table 2 we can see that TF-IDF (Term Frequency - Inverse Document Frequency) is largely used in our researched literature as a pre-processing technique. We also observe that Tokenization is not as widely used compared to TF-IDF, but it is the second most reoccurring pre-processing technique among the studies.

The third matrix (Table 3) compares traditional machine learning algorithms used for categorization in our reviewed studies. The algorithms compared are Logistic Regression (**LR**), Support Vector Machine (**SVM**), Naïve Bayes (**NB**), K-Nearest Neighbors (**KNN**), Decision Tree (**DT**), Random

Forest (**RF**), and Extreme Gradient Boosting (XGBoost/**XGB**). The symbol X indicates that the algorithm has been used in this study and the symbol O indicates that the algorithm has been used and was the best-performing one in the study. Note that not all studies have one singular best-performing algorithm and that the best-performing algorithms might not have been one of the traditional machine learning ones.

Table 3

Matrix of Literature Review: Traditional Machine Learning Algorithms

#	Reference	**Traditional Machine Learning Algorithms**						
		LR	**SVM**	**NB**	**KNN**	**DT**	**RF**	**XGB**
1	Harikrishnakumar et. al (2019)	X	O	X	X	X		
2	Chow (2022)						X	X
3	Jain (2019)			X			O	
4	Shah et al. (2020)	O			X		X	
5	Kambar et al. (2021)	X	X			O	X	X
6	Bangyal et al. (2021)	X	X	X	X	X	O	
7	Wang et al. (2019)		X				X	
8	Paduraru & Breaban (2019)	O	X			X		
9	Arkok & Zeki (2020)			X	X	X	O	
10	Ali et al. (2019)							
11	Occhipinti et al. (2022)	X	X	X	X		O	O
12	Qi (2020)		X	X	X			O
13	Taha & Yousif (2023)	X				X	X	X
14	Deng et al. (2018)		X				O	O
15	Dubey et al. (2022)	X	O		X	X	X	X
16	Gaurnav et al. (2021)	X	X					O
17	Endalie & Haile (2021)		X			X	X	X
18	Parida et al. (2021)			X			X	
19	Kamath et al. (2018)	X	X	X		X	X	
20	Suneera & Prakash (2020)	O	X	X	X	X	X	

The fourth and last matrix (Table 4) is similar to the third (Table 3) but it's not comparing traditional machine learning algorithms, instead, it is comparing different types of machine learning algorithms called Neural Networks. The algorithms compared are Multi-Layer Perceptron (**MLP**), Convolutional Neural

Network (**CNN**), Artificial Neural Network (**ANN**), and Long Short-Term Memory Network (**LSTM**). Like Table 3, the symbol X indicates that the algorithm has been used in this study and the symbol O indicates that the algorithm has been used and was the best performing one in the study. Note that not all studies have one singular best-performing algorithm and that the best-performing algorithms might not have been a Neural Network.

Table 4

Matrix of Literature Review: Neural Networks

#	Reference	Neural Networks			
		MLP	**CNN**	**ANN**	**LSTM**
1	Harikrishnakumar et. al (2019)				
2	Chow (2022)			O	
3	Jain (2019)				
4	Shah et al. (2020)				
5	Kambar et al. (2021)				
6	Bangyal et al. (2021)	X	O		X
7	Wang et al. (2019)	X	O	X	
8	Padurariu & Breaban (2019)				
9	Arkok & Zeki (2020)				
10	Ali et al. (2019)		X		
11	Occhipinti et al. (2022)				
12	Qi (2020)				
13	Taha & Yousif (2023)				
14	Deng et al. (2018)				
15	Dubey et al. (2022)			X	
16	Gaurnav et al. (2021)				
17	Endalie & Haile (2021)	X	O		
18	Parida et al. (2021)				
19	Kamath et al. (2018)		O		
20	Suneera & Prakash (2020)	X	X		X

After reviewing and comparing the 20 studies in the four matrices, we observe that TF-IDF was the most widely used pre-processing technique followed by Tokenization. We also observe that there is not a machine learning algorithm that widely performs the best for most categorization cases. Some studies have regression algorithms as the best performing, some have decision tree algorithms and some have neural network algorithms. The three algorithms that have the most reoccurring instances of performing the best are the Random Forest, XGBoost, and the Convolutional Neural Network.

2.5 Machine Learning Algorithms for Categorization

Spend Categorization

Spend categorization, is not a robustly documented topic in the literature. Nonetheless, Chow (2022) manages to classify expenses using machine learning algorithms such as XGBoost, Random Forest, and Neural Networks. Chow mentions that a balance of classes was performed (without specifying what balancing technique was used), which accounts for the fact that certain types of categories appear more frequently on the training data and corrects for it. Comparing the results obtained out of the three machine learning algorithms, Chow records that the Neural Network approach had the best performance with an accuracy of 83% when categorizing the balanced testing data. On the other hand, XGBoost had an inferior performance predicting the balanced testing data with 70% accuracy. Finally, Random Forest had the poorest performance hitting 69%. It is worth mentioning that the number of entries in this study was relatively small, with 450 entries in total for training and testing data, which should be taken into consideration while analyzing the results given.

Size of the Dataset

The sheer number of records in a particular dataset has bearing on the overall result and accuracy of the classification algorithm, as exemplified by the research performed by Ali et al. (2019). This research performed classification on seven different datasets that had similar records, but varying amounts of records. The seven datasets tested had 15k, 30k, 45k, 60k, 75k, 90k, and 100k records respectively. The results showed a clear correlation between the number of records in the dataset and a higher accuracy of the model. The results of each dataset were respectively 84.67%, 85.12%, 85.89%, 86.78%, 87.21%, 87.63%, and 88.34%.

From these results, we can observe that increasing the number of records by roughly seven times. from 15,000 records to 100,000. only increased the accuracy by 3.67%. For this particular dataset, we can conclude that even though there is a correlation between the number of records in the dataset and the accuracy of the model, most of the accuracy was obtained within the first 15,000 records.

Text Categorization

From our research, we learned that the terminology that is most widely used in the literature to describe the type of classification our sponsor company seeks to perform is *Text Categorization.* During the 1980s, the most used engine for text categorization was Knowledge Engineering, which consisted in acquiring a set of rules and categories from subject matter experts and using those to correctly categorize the imputed data. After the 1990s this industry standard began to shift more into machine learning, which offered all the benefits that Knowledge Engineering did without the need for any input from a subject matter expert. Instead, machine learning used a set of previously categorized (or labeled) data, recognized the patterns and characteristics of said categories based on the data associated with them, and did its assumptions of what would have previously been the subject matter expert input (Sebastiani 2002).

Bangyal et al. (2021) tested the accuracy of a substantially large array of different classification algorithms for text categorization. Analyzing their findings can give us an understanding of how these fare against each other under similar circumstances. The machine learning algorithms tested against each other in this study were Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbors, Adaboost, Multi-Layer Perceptron, and Naïve Bayes. The dataset classified had 10,202 records of news headlines about COVID-19, and the objective was to determine, by the words on the headline, whether the news was considered fake news or not. This number of records is relatively large compared to most of the other studies found in the literature and it’s within one order of magnitude of the number of different records we work with in our capstone project. The outcome of Bangyal et al. (2021) shows that for this particular dataset, all machine learning algorithms performed well but the best-performing ones were the Multi-Layer Perceptron, K-Nearest Neighbors, and the Random Forest all with 97% accuracy.

Another comparative study was performed by Shah et al. (2020) in which Logistic Regression, KNN, and Random Forest algorithms were tested to classify segments of a newspaper based on their topic. These topics were business, entertainment, politics, sports, and technology. This study utilized TF-IDF (which we will look into in Section 2.6) to pre-process its data. Their results show that the model with the highest accuracy for their particular dataset was Logistic Regression with an accuracy of 97%.

Wang et al. (2019) followed a similar approach in their research to categorize clinical texts such as clinical notes and progress reports. Firstly, they pre-processed their data with Natural Language Processing and subsequently ran machine learning algorithms to compare the results. The noteworthy algorithms they used were Support Vector Machine, Random Forest, and two Neural Network algorithms, namely Multi-Layer Perceptron and Convoluted Neural Networks. The results of this study showed that all algorithms performed with more than 80% accuracy. The Neural Network algorithms performed slightly better than the Random Forest.

Harikrishnakumar et. al (2019) attempted to classify the supplier base targeted by their research through a series of regression algorithms. They mainly compared different types of regressions to predict the quality of suppliers as well as other much simpler supervised algorithms such as K-Nearest Neighbors and Naïve Bayes. They did not examine any decision tree algorithms or similar ones. The best predictive accuracy for suppliers in the case of Harikrishnakumar et. al (2019) was obtained by the linear model of SVM (Support Vector Machine) with an accuracy of 87%, surpassing that of a Logistic Regression which only scored 76% accuracy. Any of the simpler supervised algorithms scored much lower with less than 60% accuracy.

Recently in 2023, Taha & Yousif (2023) performed a study comparing XGBoost, Random Forest, Decision Tree, and Logistic Regression while categorizing 127,600 records of data from news articles into four categories (World, Sport, Business, and Science/Technology). The results of this study showed a very similar accuracy among all four. The scores were 91.64%, 91.66%, 91.63%, and 91.55% respectively.

These results are virtually identical which shows that for this particular dataset, all four algorithms are good choices as a categorization technique.

A study that used machine learning to classify types of cancer as benign or malign was done by Deng et al. (2018). They compared the results between XGBoost, Random Forest, and Support Vector Machine. The results found that XGBoost and Random Forest were superior at correctly identifying the type of cancer based on a list of features of each patient.

A similar study on breast cancer was performed by Dubey et al. (2022). In it, attributes of breast lumps were utilized as features to diagnose whether the lump was benign or malign. The algorithms compared in this study were Support Vector Machine, K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost (a version of XGBoost known to excel on binary decisions), and Artificial Neural Networks. In this study, unlike the previous one, the best predictive scores were obtained by the Support Vector Machine and the Artificial Neural Networks, followed by the Logistic Regression as a close third.

Gaurnav et al. (2021) compared algorithms to identify whether a particular code was written by a human or by a “bot” (an artificial intelligence-powered machine). The algorithms applied in this study were Support Vector Machine, Logistic Regression, and XGBoost. Among these three algorithms, XGBoost performed better in terms of predictive accuracy.

Endalie & Haile (2021) collected 3,600 news samples as their dataset and categorized them into the following six categories: Business, Education, Health, Sport, Politics, and Technology. These categories were chosen arbitrarily by the author because they were the most common categories used in previous studies. To clarify the news samples, they ran the following algorithms: Support Vector Machine, Decision Tree, Random Forest, XGBoost, Convolutional Neural Networks, and Multi-Layered Perceptron. The results showed that the Convolutional Neural Network was the best-performing algorithm in terms of accuracy. All of the other algorithms performed similarly as a second-best option except the Decision Tree which greatly underperformed.

A similar study about news articles was performed by Parida et al. (2021). In this study, the author only compared the performance of Random Forest and Naïve Bayes algorithms. Also, these two algorithms were pre-processed with two different Natural Language Processing algorithms to increase the performance of the categorization. These two pre-processing algorithms were the TD-IDF (Term Frequency - Inverse Document Frequency) and the Count Vectorizer. The result in this case showed that the Random Forest performed better with the TF-IDF as a pre-processor and the Naïve Bayes performed better with the Count Vectorizer as a pre-processor. Overall, the Naïve Bayes performed slightly better in terms of accuracy.

Kamath et al. (2018) did a robust comparison between machine learning algorithms on two distinct datasets. The first one was a health dataset obtained from a private insurance company and the second one was a publicly available tobacco dataset. The machine learning techniques compared were a linear version of a Support Vector Machine, Naïve Bayes, Logistic Regression, Random Forest, Multi-Layer Perceptron, and Convolutional Neural Networks. The best-performing algorithm in terms of accuracy was the Convolutional Neural Network and a close second was the Logistic Regression.

Suneera & Prakash (2020) categorized a dataset of news documents into 20 different categories of the most common subjects of news. The set contained 18,846 records and was categorized by running three distinct pre-processing Natural Language Processing techniques as well as eleven distinct Machine Learning techniques. The pre-processing techniques utilized were TF-IDF, Word2Vec, and BERT (Bidirectional Encoder Representations from Transformers). The Machine Learning algorithms used to categorize the dataset were Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, K-Neatest Neighbors, Naïve Bayes, Multi-Layer Perceptron, Convolutional Neural Network (CNN), LSTM (Long Short-Term Memory Network), and combinations of CNN and LSTM. Out of all the possible combinations of pre-processing NLP techniques and ML techniques, the best performing was TF-IDF combined with Logistic Regression. The second-best performing was the Convolutional Neural Network.

Jain (2019) compared two algorithms, the Random Forest and the Naïve Bayes to determine which was better at predicting the sentiment of the writer when writing a movie review. The results of this study showed that, for this particular dataset, Random Forest had a significantly better performance than the Naïve Bayes.

Occhipinti et al. (2022) conducted a very exhaustive comparison between machine learning algorithms to categorize a database consisting of Enron’s emails. The goal of the algorithm was to classify the emails as spam or non-spam emails. This means that the problem was a binary-class problem. The study compared the following machine learning algorithms: Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, Random Forest, and XGBoost. The results showed that Random Forest and XGBoost were the best-performing algorithms in terms of accuracy, achieving similar results. This study also used TF-IDF as a pre-processing technique.

2.6 Pre-Processing Techniques

In most instances, the data provided to the machine learning algorithm should be enough to output complete results, nonetheless, there are ways to enhance the performance of the algorithms. Pre-processing techniques are one of these ways. Pre-processing the data with these techniques allows the machine learning models to understand the data inputted and the relations among the data’s records in a deeper, more insightful way. There are various types of pre-processing techniques. This study discusses techniques for Imbalanced Classes and Natural Language Processing (NLP).

Imbalanced Classes

The imbalance of classes occurs when certain sets of classes (or categories) are underrepresented in the data and thus, machine learning classifiers cannot properly discriminate them (Padurariu & Breaban 2019).

Padurariu and Breaban (2019) compared the two main types of method to balance classes. These are: (1) Over-sampling methods, where the algorithm creates new records of the under-represented classes, and (2) under-sampling methods, where the algorithms discard a set number of records from the overrepresented classes. The study concludes that under-sampling methods vastly underperformed compared to over-sampling methods.

Diving deeper into the different over-sampling algorithms compared in Padurariu's research, they examined (1) Random Over-sampling: where records of the under-represented classes get duplicated at random, (2) SMOTE: Synthetic Minority Over-Sampling Technique, (3) WEMOTE: a more efficient way to generate synthetic samples according to Padurariu, and (4) CWEMOTE: a version of WEMOTE that first clusters the training data of the under-represented class. The results of the experiment showed that the quality and accuracy of the algorithms improved when an over-sampling technique was implemented.

Another case of class imbalance was presented by Kambar et al. (2021), where they classified descriptive texts about drugs based on the MOA (Mechanism of Action) of each drug. In this paper, they also used SMOTE as an over-sampling method to fix the class imbalance issue.

In Arkok and Zeki (2020), an even deeper analysis of the SMOTE technique was performed. This research classified each verse of the Quran book into different classes based on the topic of the verse, and then proceeded to assess the accuracy of various machine learning classification algorithms with and without the SMOTE technique applied beforehand. The analysis of this research is quite clear; it mentions that the results of all the classification algorithms improved greatly after applying SMOTE. Without applying SMOTE, the classification algorithms had an accuracy between 55% and 75%. After applying SMOTE they had an accuracy between 77% and 96%.

In conclusion, Imbalanced Classes is a recurring topic of focus in the machine learning categorization literature, over-sampling techniques have shown to be superior to under-sampling techniques in some cases, and the SMOTE is a prevalent and commonly used technique for over-sampling.

Natural Language Processing (NLP)

Natural Language Processing (NLP) is the science that uses computational techniques to understand and generate natural language. It was first introduced to the literature in the 1940s and its first applications were on the machine translation of texts between languages (Jones 1994). One of the main uses of NLP is information extraction (Chowdary 2020, p.12) which consists of performing computational analysis on a dataset of natural language to automatically return structured data that can be queried to obtain better insights (Chowdary 2020, p.646).

NLP is comprised of an array of different techniques that are designed to mimic a human-like understanding of text (Liddy, 2001). In this study, we examine two techniques in more detail. These techniques are Tokenization and Term Frequency – Inverse Document Frequency (TF-IDF).

Tokenization

The input given to computers is often a string of characters that are then processed together as a whole. The concept of tokenization consists of dividing this string of characters into multiple smaller portions called tokens. This is often a technique performed during the earlier phases of the analysis, although not always the first one. The technique does not consist in merely splitting words apart, it also takes into account compound tokens like "in spite of" which consist of three words. One common obstacle that tokenization has is punctuation and symbols. Initially, words like "don't" would not be considered as one single token because the apostrophe symbol divides the letters apart, but the tokenization algorithm works around this to understand the underlying meaning of the string of characters (Greffenstette 1999).

Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a Natural Language Processing technique used to rate and rank the importance of a term inside a dataset. The importance of the term is calculated by multiplying the frequency of a term inside any given document (record), times the inverse of the frequency of documents (records) that contain the specific term inside a whole corpus of documents (the dataset) (Qi. 2020).

2.7 Conclusion

In this section, we reviewed the literature on the essential topics to carry out both the qualitative and quantitative components of our project. For the qualitative component, we reviewed the literature on supply chain control towers and different strategies for software implementation. For the quantitative component, we reviewed machine learning algorithms and pre-processing techniques, both in the context of text categorization. Some of the most prominent machine learning techniques for Text Categorization are Neural Networks, different variations of classification trees (Decision Tree, Random Forest, and XGBoost), and regression algorithms such as Logistic Regression. Depending on the case and the data, different algorithms may provide a better accuracy. We also discussed pre-processing techniques that could potentially enhance the performance of the machine learning model.

3 DATA AND METHODOLOGY

In this section we describe the data provided by our sponsor as input. We then discuss the methods applied to achieve the objectives of the research (mentioned in Section 1.5).

3.1 Exploratory Database Analysis

The data provided to us consisted of 713,365 records (rows of data) about the spend data of the assigned business unit in the year 2022. This database had 14 features (columns) that describe each record (row). Some features would be factored in for our category/sub-category predictions and others would not. The selection of the features was based on a qualitative analysis of whether each feature was relevant for the outcome category/sub-category. The full list of the features, the decision of whether each feature was factored in, and a rationale behind the decision for not considered features are illustrated in Table 5.

Table 5

Direct Procurement Spend Data: Features Selection

#	Feature Name	Included	Excluded	Rationale for Exclusion
1	Material Code	X		
2	Material Description	X		
3	ERP Vendor Description	X		
4	Franchise Group		X	It is correlated to Franchise.
5	Franchise	X		
6	Category	X		
7	Sub-Category	X		
8	Commodity Level 1	X		
9	Commodity Level 2		X	Insufficient data.
10	Region		X	Insufficient data.
11	Plant/Warehouse	X		
12	Plant/Warehouse Name		X	It's correlated to Plant/Warehouse
13	Plant Country Code		X	Insufficient data.
14	Spend Amount		X	Not required for categorization.

Altogether, we selected 8 features to include in our analysis. We are confident that this number of features is large enough to draw robust conclusions and that the absence of the 6 features that were disregarded won't diminish the success of our predictions.

3.2 Cleaning of the Database

The data given to us was significantly large in terms of cardinality (number of records/rows). To reduce cardinality for more efficient handling of the data, the cleaning process required a few steps. Firstly, after aligning with our sponsors, we deleted certain non-relevant records for some of the categories. Secondly, we deleted some of the Material Codes that were irrelevant from a business perspective. Thirdly, we cleaned the data to ensure that category labels were worded consistently, so the algorithm wouldn't mistakenly consider two different categories when they were the same. Finally, we deleted all duplicate records (based only on the included 8 features) and ended up reducing the cardinality of the database from 713,365 original records to 71,547 unique records.

From this database, we were also able to extract the possible categories and sub-categories of the entries that comprise our sponsor's spending. **The names of the categories and sub-categories in this study are protected for confidentiality reasons.** The following is the list of all the category names:

1. Category 1
2. Category 2
3. Category 3
4. Category 4
5. Category 5
6. Category 6
7. Uncategorized Spend

The number of records associated with each category and the percentage of each are displayed in Table 6.

Table 6

Number and Percentage of Unique Records Associated with Each Category

Category Name	Number of Records	Percentage of All Records
Category 1	32	0.04%
Category 2	610	0.85%
Category 3	12,759	17.83%
Category 4	2,802	3.92%
Category 5	26,229	36.66%
Category 6	9,301	13.00%
Uncategorized Spend	19,814	27.69%
TOTAL RECORDS	**71,547**	**100.00%**

The records of the top 6 categories of Table 6 (the ones labeled as possible outcome categories) would serve as training and testing data for the machine learning model. The records under "Uncategorized Spend" would be treated as external data records to apply the machine learning algorithms to predict their categories/sub-categories.

The database also allowed us to retrieve an exhaustive list of the 113 possible sub-categories found in Appendix A. It also indicates the count of how many records in the dataset were labeled with each sub-category and the percentage that this count represents from the total pool of labeled records.

3.3 Methodology

In the methodology section, we discuss the methods that were ultimately utilized to process the data and achieve the objectives of this research (mentioned in Section 1.5). We will also briefly discuss why some methods previously used in the literature were not included in our approach. The first part of this section will discuss the methods utilized for the qualitative component of our research and the subsequent part will discuss the artificial intelligence techniques implemented in the quantitative component of this research.

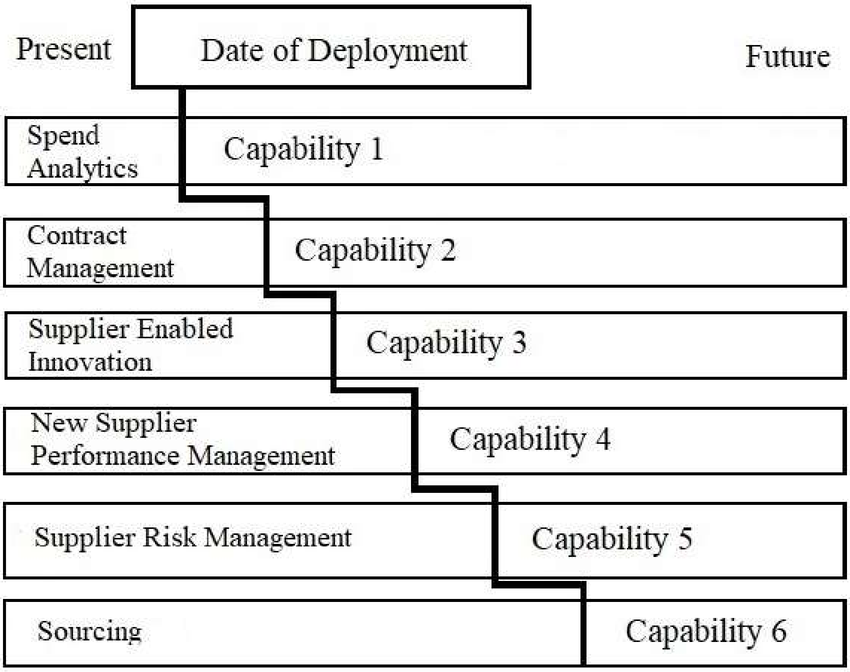
3.3.1 Procurement Control Tower Implementation Strategy

We considered various software implementation strategies. Ultimately, we selected a **Phased Implementation Strategy** as the foundational approach to build the architecture of the Procurement Control Tower. A phased approach in our specific case means that the Procurement Control Tower must be divided into sub-elements that will be deployed sequentially. The use cases/capabilities of the Procurement control tower act as the sub-elements in this case. Each capability must be able to operate individually for the phased approach to work. Only after multiple capabilities are fully functional can the interconnectivity between them be introduced.

The implementation of the phased approach is illustrated in Figure 2. On the left side of the figure, for illustrative purposes only, we have placed 6 possible use cases of the Procurement Control Tower. These examples come from the alternative use cases originally proposed in Section 1.3. The first use case of the Procurement Control Tower, Spend Analytics, is addressed by Capability 1, which is the categorization of spend data. Each use case is assigned a capability that will, in turn, provide value to the procurement area. The capabilities of the Procurement Control Tower should be deployed as a series of steps in chronological order. Each step of the phased approach should have its implementation timeline, objectives, and deliverables, similar to how this study does with Capability 1.

Figure 2

Phased Implementation Strategy of the Procurement Control Tower



Adapted from: Khanna & Arneja (2012) Choosing an Appropriate ERP Implementation Strategy. IOSR Journal of Engineering. p480

One of the main advantages of a phased approach is that it builds credibility for the overall project throughout an organization. Building this initial credibility is especially valuable in our case because our sponsor, being a large corporation, would need a high degree of conviction to implement a new system at an organizational level. Another advantage of a phased approach is that it allows for feedback and corrective actions along the steps of the implementation. This factor reduces risk because adjustments would only have to be made at a lowers scale, and not at an organizational level. A third key advantage of a phased approach is that since the implementation is spread out over time, so is the capital investment required to implement it (Hilton, 2018).

3.3.2 Logistic Regression

The first machine learning model we decided to use to categorize our sponsor's spend data was Logistic Regression. Logistic Regression is a machine learning algorithm that finds a relationship between a set number of variables through a linear equation and then utilizes it to predict the future value of one variable based on the value of the others (AWS.amazon.com, n.d). The benefit of logistic regression is that it streamlines the mathematics to find the correlation between a set of multiple variables affecting one (Lawton. G, n.d.)

If the Logistic Regression algorithm can find a linear relationship between one or more of the input variables of our dataset and the desired category and sub-categories that our model must identify, it could potentially provide good results in terms of predictive power. We chose to begin our analysis with Logistic Regression because since it is a linear model, it is one of the simplest machine learning algorithms to use. However, based on our findings of the State of the Art illustrated in Table 3, we know that in many cases the Logistic Regression algorithm captures the complexity of a problem enough to provide excellent results.

3.3.3 Decision Tree

The second machine learning categorization model that we used in our data is the Decision Tree. This model is fast to learn from the given data and predict an outcome. It works by sequentially splitting the data into at least two parts or "nodes" from top to bottom, creating a hierarchical structure with several layers that resemble a tree. Each node decides in which way to divide the data until it reaches a final node at the bottom that labels the pieces of data. The algorithm selects which attribute each node should test. It also sets the cutoff level for that particular attribute that will determine which node of the next hierarchical level the data will advance to. (Kowsari et al. 2019)

Unlike our first algorithm, Logistic Regression, the Decision Tree does not explain the relationship among the variables through a linear equation. This makes the Decision Tree our first non-linear equation implemented in our study. By being non-linear, this model can capture increased complexity from a problem that could not be otherwise captured by a linear equation.

To illustrate the work that the Decision Tree algorithm would perform with our sponsor's data, the steps it would perform would be the following:

1- Begin the tree at the highest level with one node that includes every record of the data.

2- Select one feature, and a cutoff level of this feature to split the records into at least two nodes that will be in the second level of the tree.

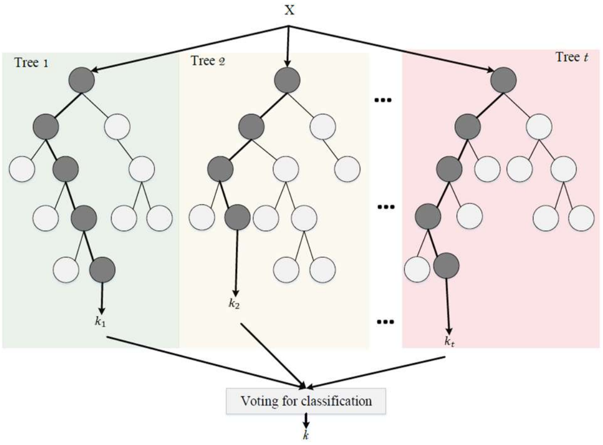
3- Repeat this same process in each of the new nodes created and continuously do so until there is enough information to perform a prediction of the category or the sub-category that should be assigned to each record.

3.3.4 Random Forest

Random Forest, just like Decision Tree, is a non-linear machine learning algorithm. It is called an ensemble technique because it combines many Decision Trees into a single model with enhanced predictive power capabilities. Its central concept is to create a finite number of randomly generated Decision Trees and then converge the predictions of all the trees into one final prediction through a voting system. Random forests train quickly on the data compared to other machine learning algorithms, but they are comparatively slow to make predictions (Kowsari et al. 2019). Figure 3 illustrates the Random Forest algorithm.

Figure 3

Random Forest Model



(Kowsari et al. 2019) Text Classification Algorithms: A Survey. p33

The x at the topmost part of the figure represents the first node where all of the data is included. This data is then processed by multiple Decision Tree algorithms. At the bottom of all the decision trees, there is an outcome categorization for each record. The categorization of a particular record might differ from tree to tree, but the algorithm selects the most frequently recurring categorization for each particular record. Because the Random Forest can use multiple Decision Trees together, it is also able to capture more complexity in a problem than the Decision Tree algorithm could.

3.3.5 Extreme Gradient Boosting (XGBoost)

The next algorithm that we ran was the Extreme Gradient Boosting or XGBoost for short. XGBoost is a model that also belongs to the Decision Tree family of algorithms just like the Random Forest. It is also an ensemble model, but the key difference is that the Random Forest uses the process of "bagging" to increase the predictive power of a Decision Tree while XGBoost uses the process of "boosting." Bagging refers to the process of generating many samples and combining the outcomes of these by a simple voting process. Boosting, on the other hand, is an iterative process that sequentially combines models to ultimately result in one best model. (Sutton, 2005) In other words, Random Forest utilizes the wisdom of the trees while XGBoost utilizes the sequential improvement of the tree.

XGBoost generates a first Decision Tree and then a second one to combine them based on a weighted decision. The algorithm places more weight on the cases which were previously misclassified to emphasize trying to correctly classify them in the next iteration of the tree. It continues to combine trees to generate a final tree that has the most robust predictive power (Sutton, 2005).

3.3.6 Natural Language Processing: Tokenization & TF-IDF

Apart from machine learning categorization models, our research also incorporated a method called Tokenization. Tokenization is a Natural Language Processing method used as a pre-processing technique that essentially breaks down text into smaller pieces called "tokens" and inputs them into the model as new data to improve the predictive power of the categorization models. (Perkins 2010)

We performed this tokenization on the Material Description feature of our database since it had the most amount of text among all the features. We also considered it to be the most important driver for a categorization decision. After the tokenization, we applied the TF-IDF (Term Frequency – Inverse Document Frequency) technique to the tokens which ranks them in terms of importance based on the TF-IDF index. The calculations of this index are explained in detail in Equations 1, 2, and 3.

$$TF - IDF\ Index = Term\ Frequency\ \times Inverse\ Document\ Frequency \tag{1}$$

$$Term\ Frequency\ = \frac{Number\ of\ occurrences\ of\ term\ inside\ a\ record}{Total\ number\ of\ words\ inside\ a\ record} \tag{2}$$

$$Inverse\ Document\ Frequency\ = \frac{Number\ of\ records\ inside\ the\ dataset}{Number\ of\ records\ containing\ the\ term} \tag{3}$$

We chose to only keep the 10,000 highest-ranked tokens based on their individual TF-IDF index. This means that each record would have one or more tokens assigned to it. These tokens would be sure to be among the 10,000 ones with the highest TF-IDF index, and the categorization machine learning models would in turn use the input of these tokens to perform a more accurate categorization.

3.3.7 Imbalanced Classes: Borderline-SMOTE

There is an imbalance of classes when a dataset has multiple different categories of data, and some of these categories have significantly fewer records than others. In these cases, the underrepresented classes are called the minority class and the overrepresented classes are called the majority class. At the same time, there might be two types of imbalances, a class imbalance, and a within-class imbalance. The first one is a simple imbalance between the main categories and the latter is an imbalance between the sub-categories enclosed by one of the main categories. (Han et al., 2005) Our research involves both types of imbalances.

If a machine learning model is trained with this imbalanced data, the algorithm would be able to learn deeply about the majority classes but not enough about the minority ones. This would produce better results outputted about the majority classes and likely underperforming results about the minority classes, creating a shortcoming in the robustness of the model.

As discussed in Section 2.4, the literature shows that the most widely used technique to address the issue of imbalanced classes is SMOTE (Synthetic Minority Oversampling Technique), which essentially creates new entries of some or all of the underrepresented classes in the dataset. This allows all classes to be similarly represented across the dataset. In our research, we will introduce a more sophisticated version of SMOTE called Borderline-SMOTE. In this context, any particular entry in a dataset could be classified into one of three options: (1) clearly belonging to a majority class, i.e., an overrepresented class, (2) clearly belonging to a minority class i.e., an underrepresented class, or (3) being close to the borderline between being classified as a majority class or a minority class. The entries on the borderline, and the ones close to it, are the hardest to correctly classify (Han et al., 2005). Borderline-SMOTE targets the entries close to the borderline on the minority side and synthetically oversamples those to achieve a stronger predictive enhancement than would otherwise be obtained by performing the traditional SMOTE.

3.3.8 Evaluation Metrics

We observed and compared the outputs of the various machine learning algorithms implemented to categorize our sponsor's spend records. According to Swalin (2018), the proper evaluation metrics for classification problems are Accuracy, Precision, Recall, and the F-1 Score. The formulas of these evaluation metrics are illustrated in Equations 4, 5, 6, and 7. To understand the underlying insights that these metrics provide we must first be familiar with four variables that form part of the formulas. These four variables are **True Positives**, **True Negatives**, **False Positives**, and **False Negatives**. The following are the definitions of these four variables with respect to our particular case:

- True Positives (**TP**): For any given category/sub-category, records that have been assigned to the category and indeed belong in it.
- True Negatives (**TN**): For any given category/sub-category, records that have **not** been assigned to it and indeed **do not** belong in it.
- False Positives (**FP**): For any given category/sub-category, records that have been assigned to it **but do not** belong in it.
- False Negatives (**FN**): For any given category/sub-category, records that **have not** been assigned to it **but do** belong in it.

With these four variables in mind, we examined the evaluation metrics of Accuracy, Precision, Recall, and the F-1 Score.

Accuracy: Our first metric, Accuracy, is a proportion of the correct predictions made by the model over the total number of predictions. We use Accuracy to measure the predictive power of our machine learning algorithms as a whole. This means that we have one single score per algorithm that reflects how well the algorithm performed when faced with the task of categorizing the uncategorized spend data. Equation 4 illustrates the index and how it is calculated (Ikonomakis et al. 2005).

$$Accuracy = \frac{TP+T}{TP+TN+FP+FN} \tag{4}$$

All of the remaining evaluation metrics (Precision, Recall, and the F-1 Score) will provide insights at a category/sub-category level.

Precision: The Precision metric reflects the percentage of records that the algorithm has correctly categorized out of all of the records assigned to any given category/sub-category. Equation 5 illustrates the index and how it is calculated (Ikonomakis et al. 2005).

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall: The Recall metric reflects the percentage of records that the algorithm has correctly categorized out of all of the records that belonged to any given category/sub-category. Equation 6 illustrates the index and how it is calculated (Ikonomakis et al. 2005).

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F-1 Score: Optimizing the Precision metric would minimize the occurrence of False Positives, and optimizing the Recall metric would minimize the occurrence of False Negatives, but optimizing for one may come at the cost of the performance of the other (Bex, 2021). For this reason, deciding to optimize for any of these two metrics will depend on whether each problem has a higher concern for False Positives or False Negatives.

Due to the nature of our problem, we have an equal degree of concern for both False Positives and False Negatives. For the types of problems that seek to optimize for both metrics simultaneously, the F-1 Score is the standard. The F-1 Score is the harmonic mean of the Precision and the Recall (Bex, 2021). Its calculation is illustrated in Equation 7.

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

3.4 Conclusion

After the cleaning of the dataset presented to us, we performed four different traditional machine learning models for categorization, a pre-processing Natural Language Processing technique called Tokenization enhanced with TF-IDF, and Borderline-SMOTE to address the imbalance of the classes in the dataset. We stopped at traditional machine learning models and did not go into Neural Networks because we assessed that the traditional models had sufficiently captured the complexity of our problem and that incorporating Neural Networks would not bring improved results to the case. To evaluate the performance of our categorization machine learning algorithms, we use the metrics of Accuracy, Precision, Recall, and F-1 Score.

4 RESULTS AND ANALYSIS

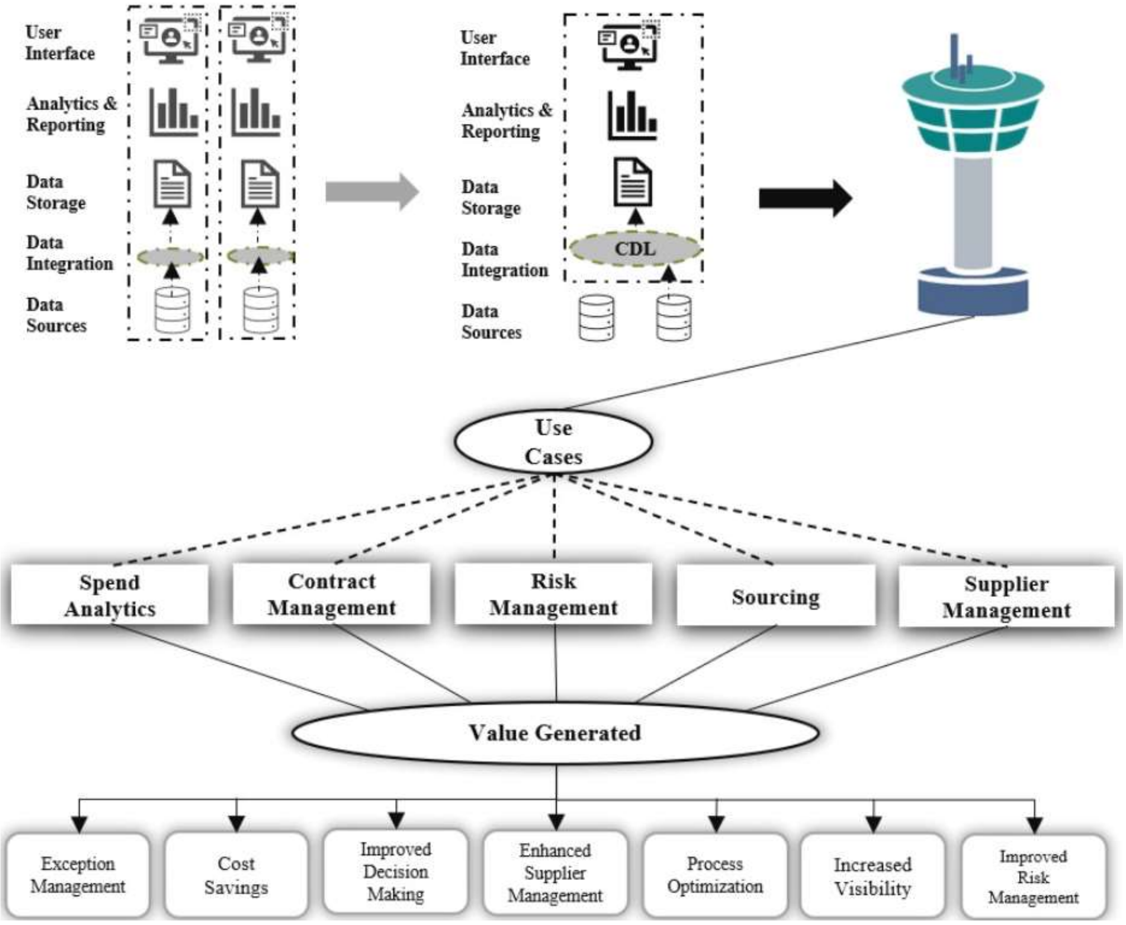
In the results and analysis section, we present the qualitative and quantitative deliverables referred to in Section 1.4. The qualitative results will include the architecture, the value proposition and the deployment strategy of the Procurement Control Tower, and the quantitative will be a comparative analysis of the different machine learning algorithms performed to show proof of concept for the prototype of the control tower's selected use-case/capability.

4.1 Qualitative Results

The qualitative results of this research explain in detail the overarching architecture of the Procurement Control Tower as well as the value proposition that it delivers. This architecture and value proposition are illustrated in Figure 4.

Figure 4

Procurement Control Tower: Architecture and Value Proposition



In essence, the Procurement Control Tower must first converge the data from our sponsor's supporting technologies/data sources into one common data layer (CDL). This first step is pivotal to have one version of the truth on which the rest of the control tower stands. This comprehensive dataset will allow the information to be stored in a single point for easy information retrieval. All of the data analysis and reporting of the pertinent areas will then be performed from a single source of information, eliminating any possible discrepancies due to a difference in the source data. Next, the user interface must be unique in the sense that every user should have the same experience, whether on a mobile or desktop device, to further bring together the users' understanding of the data. This unified data trains the Procurement Control Tower and enables it to perform enhanced data analysis. The control tower then uses the insights of this enhanced data analysis in specific use cases/capabilities. Some examples of the use cases are the alternative ones originally proposed in Section 1.3 (Contract Management, Risk Management, Sourcing, and Supplier Management). Finally, each use case/capabilities will be able to create new value such as Enhanced Exception Management, Increased Visibility, and Process Optimization.

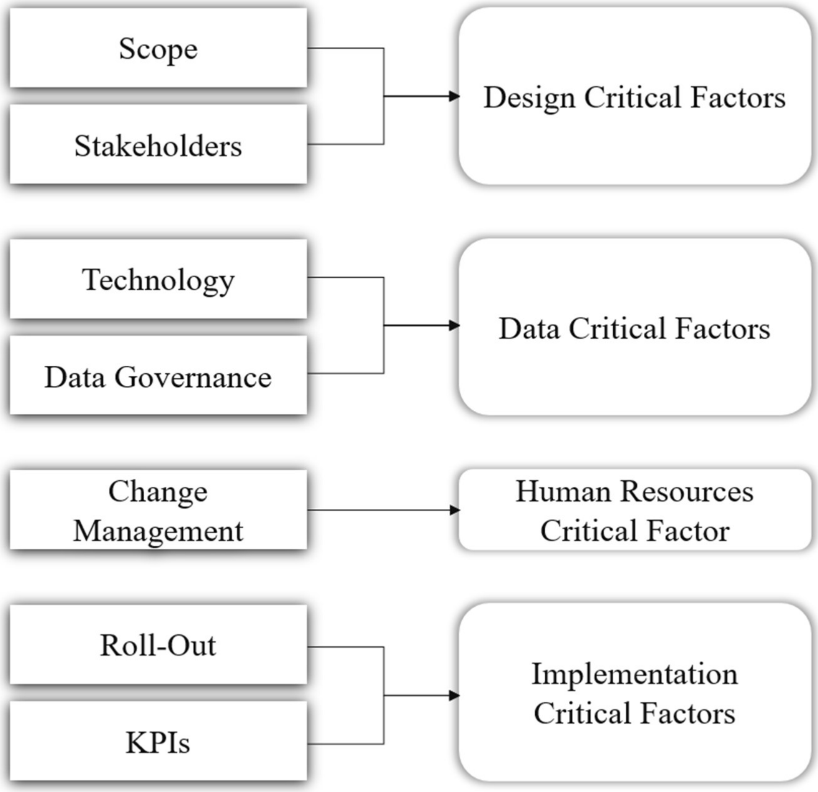
Our capstone project focused deeply on one use case, **Spend Analytics – Spend Categorization**, to thoroughly showcase the full value that can be obtained by a single capability. The benefits attained from our selected use case largely come from Improved Decision-Making, Enhanced Supplier Management, and Process Optimization. We will further analyze these findings in Section 5 (Discussion).

The development of the other use cases that the Procurement Control Tower could include in the full-scale version of the system, which our sponsor should ultimately develop, should follow a similar approach as this study has with the Spend Analytics use case. They must first pass through a development phase to later be aggregated to the control tower. This aggregation must be done as a **Phased Implementation Strategy**, as discussed in Section 3.3.1, where the use cases should be deployed in a series of steps in chronological order. The use cases should be able to function individually at the time of deployment. At a later stage the use cases can be integrated to derive the benefits of synergy. This will allow the full-scale version of the Procurement Control Tower to have the potential to generate a greater value than the sum of its parts.

Before starting the deployment process of a new system, there are critical factors that should be taken into consideration to promote a streamlined course of action. The seven critical factors that we deem to be important for the deployment strategy of the Procurement Control Tower are illustrated in Figure 5. On the left side of the figure, we observe the seven critical factors which are: Scope, Stakeholders, Technology, Data Governance, Change Management, Roll-Out, and KPIs. On the right side of the figure, we observe the type of critical factor that each one is.

Figure 5

Procurement Control Tower: Deployment Strategy Critical Factors



Narrowing down on each factor:

- **Scope:** The first critical factor that must be taken into consideration is the scope of the Procurement Control Tower. The sponsor company must formally establish the extent of what the Procurement Control Tower is expected to do and not do. This includes defining what is the full array of use cases/capabilities that the Procurement Control tower will possess and the full list of business units that the tower will overlook.
- **Stakeholders:** The sponsor must identify all the individuals, groups, and entities that have an interest or will be affected by the implementation of the Procurement Control Tower. The system must be designed in a way that is tailored to the interaction or relationship with each one. For

example, if the stakeholder is one of the users of the system, the user interface should be mindful of the requirements and needs this user might have to operate the system, and if the stakeholder is a supplier, the system should consider the agreements previously established with them such as data privacy.

- **Technology:** The sponsor must select the software that will house the digital platform of the Procurement Control Tower. This software could be the same one that is already housing the control towers of the Planning and Delivery areas. This technology must be able to combine and process the data coming from the other existing ERPs. If there is important data left out of the CDL due to software incompatibilities, it will greatly affect the value of the output of the Procurement Control Tower.
- **Data Governance:** Data Governance refers to the management of the data itself. It involves a set of policies and guidelines that ensure the availability, usability, and security of databases. Having a Data Governance structure places accountability on a data management team to ensure periodic data quality checks and policy compliance.
- **Change Management:** The implementation of the Procurement Control Tower must follow a deliberate Change Management strategy. This strategy should incorporate both a top-down and a bottom-up approach. By doing so, we maximize the probability of the stakeholders inside the organization being committed to the implementation of the new system. For the top-down approach, senior management must be aware and in line with the overall value that the Procurement Control Tower provides. For the bottom-up approach, the actual end users of the interface must be aware of why they are undergoing this change in their system and also asked for input when making systemic decisions of the overall control tower.
- **Roll-Out:** The sponsor company must establish the mechanism that they will use to deploy the full-scale Procurement Control Tower. We recommend utilizing a Minimum Viable Product (MVP) strategy with a phased approach to quickly demonstrate the value that the full-scale system provides and thus, gain credibility across the organization. The sponsor should also execute a

testing phase to reduce the risk due to early malfunctions, validate the systems capabilities, and ensure user acceptance.

- **KPIs:** Prior to the deployment phase of the system, the sponsor should establish the KPIs that will measure the success of the Procurement Control Tower. These will also act as one of the main sources of feedback to improve the system.

4.2 Quantitative Results

In this section, we will first present, analyze, and compare the results obtained by the different machine learning models we ran to select the best-performing model. Then we narrow down to the selected algorithm and conduct a deeper analysis of its performance at a category and sub-category level.

Accuracy

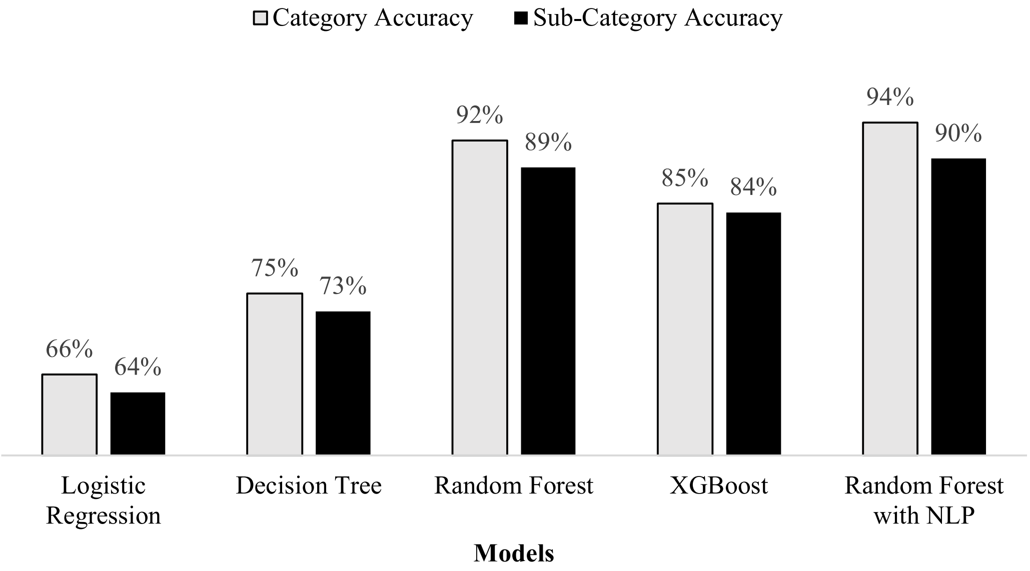
Firstly, we will present the performance of all the machine learning models tested by comparing their accuracy. This measure will indicate which are the best models to predict the categories and sub-categories of our sponsor's spend data.

In Figure 6 we can observe the percentage of records that each machine learning model could correctly classify, both at a category and a sub-category level. Our linear model, Logistic Regression, was vastly inferior in terms of predictive power compared to the non-linear models we ran. Out of our three non-linear models (Decision Tree, Random Forest, and XGBoost), the ensemble models (Random Forest and XGBoost) performed significantly better. Out of both ensemble models, Random Forest had the highest accuracy score on both the category and the sub-category level.

At this stage of the study, after testing the traditional linear, non-linear, and ensemble machine learning models, we were able to achieve results of 92% accuracy at a category level and 89% at a sub-category level. After observing the highest accuracy using Random Forest, we proceeded to enhance the quality and the predictive power of the model by pre-processing the data with NLP and addressing the imbalance of classes with Borderline-SMOTE. Since Borderline-SMOTE generates synthetic samples of minority classes, it will improve the accuracy of those. Before applying Borderline-SMOTE to the Random Forest algorithm, the results showed 16 sub-categories with 0% Precision and 0% Recall. After applying Borderline-SMOTE that number was reduced almost by half to 9. The application of NLP and Borderline-SMOTE increased the accuracy of the Random Forest algorithm both at a category and a subcategory level. The enhanced Random Forest obtained a final accuracy of 94% for categories and 90% for sub-categories.

Figure 6

Results: Accuracy of the Machine Learning Models.



Precision, Recall, and F-1 Score

To further analyze the results of our best-performing machine learning algorithm, Random Forest with NLP and Borderline-SMOTE, we will showcase the performance of the model at a category and sub-category level. We will do so by comparing its Precision, Recall, and F-1 scores in Figure 7.

Figure 7

Results: Precision, Recall, and F-1 Score at a Category Level

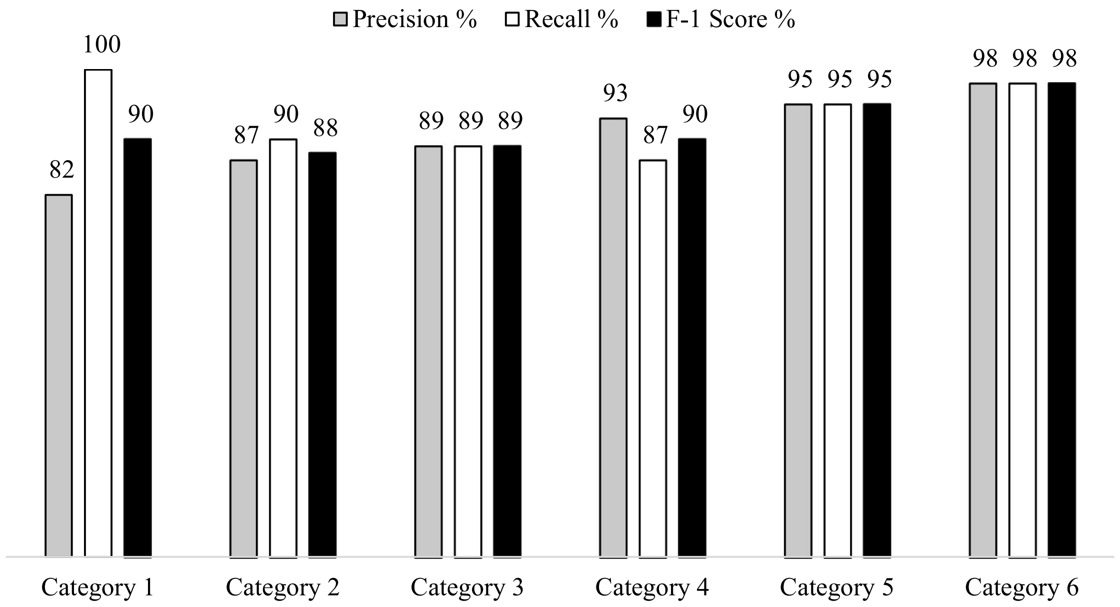


Figure 7 shows that Category 6 and Category 5 had the best results overall due to their F1 score. At the same time, we see that all categories had relatively high F1 scores, 88% being the lowest. The Precision, Recall, and F-1 score results at a sub-category level are displayed in Appendix B. Some sub-categories exist in more than one category. For example, “3A” is a sub-category that can be found in both Category 3 and Category 5. The *List of Categories and Sub-Categories* (Appendix A) enumerates both possibilities separately, but the *Results: Precision, Recall, and F-1 Score of Sub-Categories* (Appendix B) provides only one set of evaluation metrics for the sub-category “3A.” This unique set reflects the performance of the sub-category as a standalone item.

In conclusion, the algorithms used were able to capture the complexity that our problem entails. The next stage in our research would have been to test **Neural Network** algorithms since they are generally capable of capturing higher levels of complexity than traditional machine learning algorithms, but since the complexity of our problem was sufficiently captured, we decided to stop running additional models.

5 DISCUSSION

The Random Forest with Natural Language Processing and Borderline-SMOTE was the best-performing algorithm in our research and thus, the one we present to our sponsor. This is congruent with our conclusions from the literature review in Section 2.6. In that section, we presented evidence that across the current literature about the *Text Categorization* topic, the Random Forest was among the models that were often the best performing.

From the results in Section 4.2.2, we see that the algorithm has a 94% accuracy at correctly categorizing our sponsor's spend data and a 90% accuracy at correctly sub-categorizing it. This level of future insight will positively impact the business of our sponsor in a multitude of ways. By having accurate spend analytics, the company can have enhanced monitoring of purchasing behavior. This will allow the company to track and measure the volumes spent on individual categories/sub-categories, and recognize opportunities such as re-negotiation with suppliers or re-sourcing to capitalize on economies of scale in case of a growing trend. Another benefit of enhanced spend analytics is budgeting. Without accurate spend analytics as a foundation, creating and following a budget for future spending will be an increasingly difficult task. By keeping accurate spend analytics the company can leverage the value of following a multi-class budget. One last advantage offered by our model is the reduction in man-hours previously spent manually categorizing the spend data. The automation of the categorization allows for the subject matter experts to focus more on knowledge work and less on repetitive tasks that offer lower value to the company. A real example of such a repetitive task is the introduction of hundreds of new Material Codes to the spend database on a quarterly basis. Our algorithm will automatically categorize these new materials and save the man-hours that it would have otherwise taken to do so.

To look more closely into the business insights of our results, we confronted the F-1 scores at a category level with the weight of each category. The weight of the categories was determined by the percentage of the records that each category amounted from the original dataset about the 2022 spend data. The comparison is shown in Table 7. The goal of this comparison is to analyze whether the algorithm is accurately categorizing the most relevant categories. From this we can determine that the weighted average of the F-1 scores is **92%**. This weighted F-1 score reflects the accuracy of the algorithm taking into account both Precision and Recall and giving more importance to the categories that reoccur more frequently in the original 2022 database.

Table 7

Insights of Model Categorization

Category Name	F-1 Score	% of Records in 2022 Data
Category 1	90%	0.01%
Category 2	88%	0.22%
Category 3	89%	63.94%
Category 4	90%	1.38%
Category 5	95%	17.95%
Category 6	98%	16.49%
Weighted F-1 Score: 92%		

Likewise, we confronted the F-1 scores at a sub-category level with the weight of each subcategory. The weight of the sub-categories was determined by the percentage of the records that each sub-category amounted to from the original dataset about the 2022 spend data. Even though there were ultimately 93 possible sub-categories in the report, as shown in Appendix B, the 10 sub-categories reoccurring most frequently in the 2022 data accounted for 91% of the records in the entire dataset. These 10 are shown in Table 8 together with their respective F-1 score. The weighted F-1 score of the most significant sub-categories was **93%**. This means that, on average, 91% of the records will be sub-categorized with an F-1 score of **93%**. Another notable observation is that the algorithm had an F-1 score of 93% in the most reoccurring sub-category "3Q," which amounted to 41% of the total records of the original dataset.

Table 8

Insights of Model Sub-Categorization

#	Sub-Category Name	F-1 Score	% of Records in 2022 Data
1	3Q	93%	41%
2	5F	91%	14%
3	3K	91%	9%
4	3J	89%	7%
5	6AA	100%	7%
6	6O	100%	3%
7	3T	86%	3%
8	6K	94%	2%
9	3B	97%	2%
10	3L	85%	1%
Weighted F-1 Score: 93%			

After we obtained 94% accuracy at a category level and 90% accuracy at a sub-category level, **we began working in collaboration with our sponsor's Data Science team to implement our algorithm in its systems.** This prompt action by our sponsor is the first step to creating the foundation of the full-scale Procurement Control Tower as explained in Section 1.4. This timely implementation of our algorithm means that it will have an immediate impact on our sponsor's procurement strategy, decision-making capabilities, and overall stakeholder satisfaction, which will later positively impact the company's bottom line through potential cost savings. Once the algorithm is implemented in the company's system and delivers the initial insights, it is expected to run periodically as an in-house solution. This continued delivery of value is unlike previous attempts to categorize the spend data by other services contracted by our sponsor, which only categorized the data as a one-time activity.

Looking back to our research questions and objectives in this project (Sections 1.2 and 1.4), we sought to demonstrate, in a measurable way, the value that a Procurement Control Tower could create for our sponsor by introducing a prototype of one of its use cases. This prototype would in turn serve as a foundation and/or baseline for our sponsor company to follow in its effort to optimize, innovate, and streamline its procurement processes. In our Model Results Section (4.2.2) we displayed the impact that implementing one use case would have on the quality of the Spend Analytics. Now, in this Section (5), we discussed the magnitude of the business impact it would have. Due to this sizable magnitude of business impact, our sponsor decided to implement our deliverables in an immediate fashion.

6 CONCLUSION & RECOMMENDATIONS

We accepted our sponsor's invitation to take part in the design of a tool that could ultimately be the prototype of a control tower responsible for the procurement of more than 10 billion USD annually. We then delivered a working model and actionable items that generate measurable value. This value is not only generated in a singular instance, but also periodically, since the model will continue to be applied on future new data. The final algorithm of the model for our assigned business unit was presented to the sponsor, who is set to replicate the algorithm in the remaining business units.

This study utilized the power of Machine Learning and Natural Language Processing, two subsets of artificial intelligence, to add value to the core processes of a multibillion-dollar company. We used linear, non-linear, and ensemble non-linear machine learning algorithms to conduct the study. The results of the algorithms would imply that there is measurable value in implementing a control tower for the procurement functions of our sponsor company. This indicates that our hypothesis, stated in Section 1.4, was correct.

To the literature, we contributed a multi-class comparative analysis of the performance of multiple machine learning algorithms in the text categorization of spend data. We then successfully integrated Natural Language Processing with our Machine Learning model and enhanced its predictive power. The specific combination of methods that produced the best results in our study was the Random Forest algorithm, balanced with Borderline-SMOTE, and using Tokenization and TF-IDF as Natural Language Processing.

For future research, we suggest expanding the scope of capabilities to include more use cases of a Procurement Control Tower than the one we examined in this study. Doing so will support the advancement of the literature on Procurement Control Towers and ultimately may allow for a meta-analysis, one of the highest levels of evidence in academic literature, to be conducted on this topic.

Through this research, we have provided an example of how to develop a capability of a Procurement Control Tower. This example can be reproduced or used as a model for future capabilities. Our main finding, which answered our research questions (Section 1.2), was that there is measurable value in implementing a Procurement Control Tower to our sponsor's business. This value is attained by progressively equipping the Procurement Control Tower with capabilities, such as Spend Analytics, that will work individually and synergistically to add value to the business. Taking this into account, our recommendation to our sponsor is to appoint a team to lead the development of the Procurement Control Tower. The first objective of said team should be to oversee the replication of the prototype we have developed for our assigned business unit to the remaining business units of the company. This will lead to the improvement of the categorization of more than 10 billion USD that our sponsor procures annually. The second objective of the appointed team should be to decide the order in which the other capabilities of the control tower should be developed to promote a harmonious fit and a collaborative attribute among them.

Our final reflection is that the study of Procurement Control Towers is a promising subtopic of the academic literature, with potential for future investigation. This study contributes to it, and we look forward to seeing this subtopic further advance as the topics related to Supply Chain Management continue to gain recognition in the business and academic sectors.

REFERENCES

Ali, M. N. Y., Sarowar, M. G., Rahman, M. L., Chaki, J., Dey, N., & Tavares, J. M. R. (2019). Adam deep learning with SOM for human sentiment classification. *International Journal of Ambient Computing and Intelligence (IJACI)*, *10*(3), 92-116.

Arkok, B., & Zeki, A. (2020). Classification of Holy Quran Verses based on Imbalanced Learning. *International Journal on Islamic Applications in Computer in Science and Technology*, *8*, 11-24.

AWS.amazon.com (n.d.) What is Logistic Regression? https://aws.amazon.com/what-is/logistic-regression/

Bex T. (2021) Comprehensive Guide to Multiclass Classification Metrics. https://towardsdatascience.com/comprehensive-guide-on-multiclass-classification-metrics-af94cfb83fbd

Bangyal, W. H., Qasim, R., Rehman, N. U., Ahmad, Z., Dar, H., Rukhsar, L., ... & Ahmad, J. (2021). Detection of fake news text classification on COVID-19 using deep learning approaches. *Computational and mathematical methods in medicine*, *2021*, 1-14.

Chow, E. (2022) Categorizing Short Text Descriptions: Machine Learning or Not? https://towardsdatascience.com/categorising-short-text-descriptions-machine-learning-or-not-d3ec8de8c40

Chowdary, C.K. (2020) Fundamentals of Artificial Intelligence. Springer.

Davenport, T.H. (1998) Putting the Enterprise into the Enterprise System. *Harvard Business Review. August-July 1998*

Deng, X., Luo, Y., & Wang, C. (2018, November). Analysis of risk factors for cervical cancer based on machine learning methods. In *2018 5th IEEE international conference on cloud computing and intelligence systems (CCIS)* (pp. 631-635). IEEE.

Dubey, C., Shukla, N., Kumar, D., Singh, A. K., & Dwivedi, V. K. (2022, November). Breast Cancer Modeling and Prediction Combining Machine Learning and Artificial Neural Network Approaches. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 119-124). IEEE.

Endalie, D., & Haile, G. (2021). Automated Amharic news categorization using deep learning models. *Computational Intelligence and Neuroscience*, *2021*.

Gaurav, V., Singh, S., Srivastava, A., & Shidnal, S. (2022). Codescan: A supervised machine learning approach to open-source code bot detection. In *Applied Information Processing Systems: Proceedings of ICCET 2021* (pp. 381-389). Springer Singapore.

Gupta, A. (2022). What Is a Supply Chain Control Tower — And What's Needed to Deploy One? https://www.gartner.com/en/articles/what-is-a-supply-chain-control-tower-and-what-s-needed-to-deploy-one

Grefenstette, G. (1999). Tokenization. *Syntactic Wordclass Tagging*, 117-133.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1* (pp. 878-887). Springer Berlin Heidelberg.

Harikrishnakumar, R., Dand, A., Nannapaneni, S., & Krishnan, K. (2019, December). Supervised machine learning approach for effective supplier classification. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (pp. 240-245). IEEE.

Hilton, K. (2018). Considering a Phased Approach to Implementing Shared Services? https://www.scottmadden.com/content/uploads/2018/08/ScottMadden_Considering_Phased_Approach_for_Shared_Services_2018_0807.pdf

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, *4*(8), 966-974.

Jain, R. (2019). Comparative Study of Machine Learning Algorithms for Document Classification. *International Journal of Computer Sciences and Engineering IJCSE*.

Jones, K. S. (1994). Natural language processing: a historical review. *Current issues in computational linguistics: in honour of Don Walker*, 3-16.

Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018, August). Comparative study between traditional machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering 2018* (pp. 1-11).

Kambar, M. E. Z. N., Nahed, P., Cacho, J. R. F., Lee, G., Cummings, J., & Taghva, K. (2022). Clinical text classification of alzheimer's drugs' mechanism of action. In *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 1* (pp. 513-521). Springer Singapore.

Khanna, K., & Arneja, G. P. (2012). Choosing an appropriate ERP implementation strategy. *IOSR journal of Engineering*, *2*(3), 478-483.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Lawton, G. (n.d.). Logistic Regression. https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

Liddy, E. D. (2001). Natural language processing.

Occhipinti, A., Rogers, L., & Angione, C. (2022). A pipeline and comparative study of 12 machine learning models for text classification. *Expert Systems with Applications*, *201*, 117193.

Padurariu, C., & Breaban, M. E. (2019). Dealing with data imbalance in text classification. *Procedia Computer Science*, *159*, 736-745.

Parida, U., Nayak, M., & Nayak, A. K. (2021, January). News text categorization using random forest and naïve bayes. In *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)* (pp. 1-4). IEEE.

Perkins, J. (2010). *Python text processing with NLTK 2.0 cookbook* (Vol. 9). Birmingham: PACKT publishing.

Qi, Z. (2020, June). The text classification of theft crime based on TF-IDF and XGBoost model. In *2020 IEEE International conference on artificial intelligence and computer applications (ICAICA)* (pp. 1241-1246). IEEE.

SAP.com. (n.d.). Supply Chain Control Towers: Providing End-to-End Visibility. https://www.sap.com/insights/supply-chain-control-tower.html

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, *5*, 1-16.

Suneera, C. M., & Prakash, J. (2020, December). Performance analysis of machine learning and deep learning models for text classification. In *2020 IEEE 17th India Council International Conference (INDICON)* (pp. 1-6). IEEE.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics*, *24*, 303-329.

Swalin, A (2018). *Choosing the Right Metric for Evaluating Machine Learning Models - Part 2. h*ttps://medium.com/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428

Taha, W. A., & Yousif, S. A. (2023, February). Enhancement of text categorization results via an ensemble learning technique. In *AIP Conference Proceedings* (Vol. 2457, No. 1, p. 040011). AIP Publishing LLC.

Verwijmeren, M. (2017). 5 Essential Elements of the Modern Control Tower. https://www.mpo.com/blog/5-essential-elements-of-the-modern-control-tower

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., ... & Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making*, *19*, 1-13.

APPENDIX

Appendix A

List of Categories and Sub-Categories – Training Data

<table>
<tr><th colspan="4">Category 1</th></tr>
<tr><th>#</th><th>Sub-Category Name</th><th>Count</th><th>%</th></tr>
<tr><td><b>1</b></td><td>1A</td><td>28</td><td>0.06%</td></tr>
<tr><td><b>2</b></td><td>1B</td><td>1</td><td>0.00%</td></tr>
<tr><td><b>3</b></td><td>1C</td><td>3</td><td>0.01%</td></tr>
<tr><th colspan="4">Category 2</th></tr>
<tr><th>#</th><th>Sub-Category Name</th><th>Count</th><th>%</th></tr>
<tr><td><b>4</b></td><td>2A</td><td>56</td><td>0.12%</td></tr>
<tr><td><b>5</b></td><td>2B</td><td>17</td><td>0.04%</td></tr>
<tr><td><b>6</b></td><td>2C</td><td>46</td><td>0.10%</td></tr>
<tr><td><b>7</b></td><td>2D</td><td>46</td><td>0.10%</td></tr>
<tr><td><b>8</b></td><td>2E</td><td>42</td><td>0.09%</td></tr>
<tr><td><b>9</b></td><td>2F</td><td>1</td><td>0.00%</td></tr>
<tr><td><b>10</b></td><td>2G</td><td>2</td><td>0.00%</td></tr>
<tr><td><b>11</b></td><td>2H</td><td>4</td><td>0.01%</td></tr>
<tr><td><b>12</b></td><td>2I</td><td>1</td><td>0.00%</td></tr>
<tr><td><b>13</b></td><td>2J</td><td>5</td><td>0.01%</td></tr>
<tr><td><b>14</b></td><td>2K</td><td>5</td><td>0.01%</td></tr>
<tr><td><b>15</b></td><td>2L</td><td>7</td><td>0.02%</td></tr>
<tr><td><b>16</b></td><td>2M</td><td>84</td><td>0.19%</td></tr>
<tr><td><b>17</b></td><td>2N</td><td>3</td><td>0.01%</td></tr>
<tr><td><b>18</b></td><td>2O</td><td>9</td><td>0.02%</td></tr>
<tr><th colspan="4">Category 3</th></tr>
<tr><th>#</th><th>Sub-Category Name</th><th>Count</th><th>%</th></tr>
<tr><td><b>19</b></td><td>3A</td><td>1</td><td>0.00%</td></tr>
<tr><td><b>20</b></td><td>3B</td><td>26</td><td>0.06%</td></tr>
<tr><td><b>21</b></td><td>3C</td><td>63</td><td>0.14%</td></tr>
<tr><td><b>22</b></td><td>3D</td><td>4</td><td>0.01%</td></tr>
<tr><td><b>23</b></td><td>3E</td><td>217</td><td>0.48%</td></tr>
</table>

24	3F	78	0.17%
25	3G	64	0.14%
26	3H	325	0.72%
27	3I	3	0.01%
28	3J	800	1.78%
29	3K	3,896	8.65%
30	3L	323	0.72%
31	3M	26	0.06%
32	3N	187	0.42%
33	3O	102	0.23%
34	3P	247	0.55%
35	3Q	1,242	2.76%
36	3R	483	1.07%
37	3S	92	0.20%
38	3T	780	1.73%
Category 4			
#	**Sub-Category Name**	**Count**	**%**
39	4A	10	0.02%
40	4B	20	0.04%
41	4C	14	0.03%
42	4D	42	0.09%
43	4E	6	0.01%
44	4F	218	0.48%
45	4G	4	0.01%
46	4H	3	0.01%
47	4I	3	0.01%
48	4J	29	0.06%
49	4K	2	0.00%
50	4L	15	0.03%
51	4M	559	1.24%
52	4N	26	0.06%
53	4O	88	0.20%
54	4P	23	0.05%

55	4Q	8	0.02%
56	4R	37	0.08%
57	4S	36	0.08%
58	4T	8	0.02%
59	4U	61	0.14%
60	4V	29	0.06%
61	4W	618	1.37%
62	4X	27	0.06%
63	4Y	140	0.31%
64	4Z	13	0.03%
65	4AA	1	0.00%
66	4AB	3	0.01%
67	4AC	3	0.01%
68	4AD	4	0.01%
69	4AE	2	0.00%
70	4AF	82	0.18%
71	4AG	4	0.01%
72	4AH	10	0.02%
73	4AI	11	0.02%
Category 5			
#	**Sub-Category**	**Count**	**%**
74	3A	2,235	4.96%
75	3B	7,767	17.24%
76	3T	1,535	3.41%
77	5A	51	0.11%
78	5B	1	0.00%
79	5C	258	0.57%
80	5D	106	0.24%
81	5E	232	0.51%
82	5F	12,664	28.11%
Category 6			
#	**Sub-Category**	**Count**	**%**
83	3T	24	0.05%

84	6A	4	0.01%
85	6B	4	0.01%
86	6C	1	0.00%
87	6D	175	0.39%
88	6E	5	0.01%
89	6F	168	0.37%
90	6G	6	0.01%
91	6H	11	0.02%
92	6I	1	0.00%
93	6J	4	0.01%
94	6K	3,556	7.89%
95	6L	286	0.63%
96	6M	2	0.00%
97	6N	2	0.00%
98	6O	203	0.45%
99	6P	103	0.23%
100	6Q	73	0.16%
101	6R	48	0.11%
102	6S	431	0.96%
103	6T	989	2.20%
104	6U	84	0.19%
105	6V	644	1.43%
106	6W	1,084	2.41%
107	6X	18	0.04%
108	6Y	8	0.02%
109	6Z	5	0.01%
110	6AA	348	0.77%
111	6AB	2	0.00%
112	6AC	10	0.02%
113	6AD	425	0.94%

Appendix B

Results: Precision, Recall, and F-1 Score of Sub-Categories

#	Sub-Category Name	Precision	Recall	F1-Score
1	1A	100%	93%	97%
2	1C	50%	100%	67%
3	2A	90%	56%	69%
4	2B	89%	100%	94%
5	2C	100%	88%	94%
6	2D	53%	89%	67%
7	2E	100%	100%	100%
8	2H	100%	100%	100%
9	2J	100%	100%	100%
10	2L	0%	0%	0%
11	2M	96%	88%	92%
12	2O	0%	0%	0%
13	3A	93%	96%	95%
14	3B	97%	96%	97%
15	3C	84%	89%	86%
16	3D	50%	100%	67%
17	3E	96%	97%	96%
18	3F	50%	62%	55%
19	3G	100%	100%	100%
20	3H	88%	83%	85%
21	3I	100%	50%	67%
22	3J	89%	90%	89%
23	3K	91%	91%	91%
24	3L	84%	85%	85%
25	3M	14%	25%	18%
26	3N	91%	50%	65%
27	3O	71%	66%	68%
28	3P	89%	86%	88%
29	3Q	95%	91%	93%

30	3R	69%	82%	75%
31	3S	89%	86%	87%
32	3T	84%	88%	86%
33	4A	100%	75%	86%
34	4AB	0%	0%	0%
35	4AC	100%	100%	100%
36	4AD	0%	0%	0%
37	4AF	87%	91%	89%
38	4AG	100%	100%	100%
39	4AH	0%	0%	0%
40	4AI	0%	0%	0%
41	4B	25%	100%	40%
42	4C	100%	100%	100%
43	4D	100%	100%	100%
44	4E	100%	100%	100%
45	4F	100%	98%	99%
46	4I	100%	100%	100%
47	4J	78%	100%	88%
48	4L	100%	100%	100%
49	4M	96%	86%	91%
50	4N	100%	73%	85%
51	4O	86%	79%	83%
52	4P	78%	100%	88%
53	4Q	40%	100%	57%
54	4R	100%	75%	86%
55	4S	54%	100%	70%
56	4T	100%	100%	100%
57	4U	81%	94%	87%
58	4V	38%	75%	50%
59	4W	94%	90%	92%
60	4X	80%	100%	89%
61	4Y	85%	92%	88%
62	4Z	100%	100%	100%

63	5A	58%	50%	54%
64	5C	74%	96%	83%
65	5D	78%	76%	77%
66	5E	79%	10%	17%
67	5F	85%	98%	91%
68	6A	50%	100%	67%
69	6AA	100%	99%	100%
70	6AC	100%	75%	86%
71	6AD	94%	94%	94%
72	6B	100%	100%	100%
73	6D	96%	96%	96%
74	6E	0%	0%	0%
75	6F	94%	91%	92%
76	6G	100%	100%	100%
77	6H	67%	100%	80%
78	6J	0%	0%	0%
79	6K	92%	97%	94%
80	6L	91%	96%	94%
81	6N	100%	100%	100%
82	6O	85%	92%	89%
83	6P	94%	65%	77%
84	6Q	96%	100%	98%
85	6R	83%	88%	86%
86	6S	97%	92%	95%
87	6T	99%	85%	91%
88	6U	86%	95%	90%
89	6V	95%	96%	96%
90	6W	97%	98%	98%
91	6X	67%	100%	80%
92	6Y	100%	100%	100%
93	6Z	0%	0%	0%