Enabling Proactive Quality in Commercial Airplanes using Natural Language Processing

by

Christian Allinson

B.S., Electrical Engineering
University of California, Davis, 2013

Submitted to the MIT Sloan School of Management and
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Master of Business Administration

and

Master of Science in Electrical Engineering and Computer Science

in conjunction with the Leaders for Global Operations program
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2022

Author .....
MIT Sloan School of Management and
Department of Electrical Engineering and Computer Science
May 6, 2022

Certified by .....
Steven Spear
Senior Lecturer of System Dynamics
Thesis Supervisor

Certified by .....
Duane Boning
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Maura Herson
Assistant Dean, MBA Program
MIT Sloan School of Management

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Enabling Proactive Quality in Commercial Airplanes using Natural Language Processing

by

Christian Allinson

Submitted to the MIT Sloan School of Management and
Department of Electrical Engineering and Computer Science
on May 6, 2022, in partial fulfillment of the
requirements for the degrees of
Master of Business Administration
and
Master of Science in Electrical Engineering and Computer Science
in conjunction with the Leaders for Global Operations program

Abstract

Quality management systems traditionally draw insight from structured, often numerical, sources of data; unstructured, free-text representations of quality data are less frequently employed despite having high informational value, and often require additional human effort to prepare their contents for use. An ability to extract and proactively employ this information enables a richer analysis of quality performance.

The primarily free-text reports generated by Boeing Commercial Airplane's "in-service investigation" (ISI) process are taken as an example of such quality data. We investigate both an unsupervised clustering method and a supervised classification method to group these reports by the broader "quality topic" they pertain to, using semantic relationship-maintaining text "embeddings" as features. We find success in supervised classification, and describe a method to relate ISI records with quality records from other parts of the commercial airplane value stream via standardized "code" metadata.

We extend the use of similarity techniques to investigation execution and propose a "helper" tool that automates parts of the manual data collection and relationship-finding process. The benefits of using such a tool over traditional keyword searches are described through an illustrated example.

Thesis Supervisor: Steven Spear
Title: Senior Lecturer of System Dynamics

Thesis Supervisor: Duane Boning
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would first like to thank Boeing and the BCA Quality Fleet Integration team for hosting me during this internship, and for sharing so openly and inclusively. I owe particular gratitude to Meghan Wright as my supervisor, Chelsea Zarnowski and Hector Silva as my sponsors, and the larger LGO community within Boeing for their help and friendship.

I would also like to thank Duane Boning and Steve Spear for advising me throughout this project and in the thesis writing process. Their guidance helped me find the intersection of technical and organizational forces unique to what LGO teaches, and moreso helped me craft that experience into a readable thesis.

Broad-reaching thanks goes to those who have helped me get to and get through LGO. Chevron looms large and I hold experiences from ETC, Questa, and SJV dearly, especially because of the mentorship of Lisa Brenskelle, Brandon Janak, KP Ong, and Ghassan Sammour. Within LGO, I owe the best parts of my time at MIT to Team 2 (Steph, Jeff, Luke, and Kunal) and the full Class of '22.

Most importantly, I would like to thank my parents Paul and Maureen, and brothers Gregory and Jeffrey. My father has been my strongest example of technical prowess and thoughtful leadership, and my mother the same in kindness and selflessness. I hope a little bit of their character shows in this work.

Disclaimer

This thesis discuses quality management within the commercial aviation industry. All examples of quality records or quality areas are constructed for illustrative purposes only and do not represent any real occurrence.

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

The value of quality and the pursuit of quality improvements are essential to success in many businesses. Obtaining benefit from these improvement efforts is dependent on both data and intentional, proactive action; challenges often arise when valuable information is stored in unstructured forms, or when this information is only acted upon reactively. This thesis explores how business processes, particularly those supporting quality, may be improved when information from such data can be efficiently and consistently employed. Modern natural language processing and machine learning methods are explored to aid in extracting this information, with Boeing Commercial Airplanes' "in-service investigation" process used as an example. This chapter presents an overview of the value of quality and a comparison of "reactive" and "proactive" work processes, the motivation to explore this particular use case at Boeing, examples of beneficial employment of free-text data in similar contexts, and concludes with a project overview and outline of the remaining thesis chapters.

1.1 Quality & Quality Management

The concept of quality is central to many organizations, for-profit and philanthropic alike, with improving quality often found as a strategic objective. For businesses, particularly those that manufacture products, quality largely brings two major outcomes: ability to acquire and retain customers, and better financial performance through

raised efficiency and lowered costs.

In a competitive marketplace, potential customers are free to choose from an assortment of products - it is incumbent on the supplier of these products to attract customers to their offering through differentiators, such as features or cost. Consumers often cite quality as one of the most important factors they consider when choosing a product. A survey of car buyers list "quality" and "how well-made" as the two most important characteristics in an automobile [40]; just over half of general consumers report that quality is the most important factor when making a purchase, beating "price" by 15% [20].

The role of quality in manufacturing impacts the cost incurred to build a product. A simple examples is rework: a part that is built with sub-standard quality will need to be replaced with a correctly built part. This incorrectly built part not only wastes materials at some cost, but also wastes time - with a fixed workforce, each reworked part takes away from the time available to build correct, sellable product. Consistently poor build quality also raises the risk of bad product being mistakenly delivered to customers, damaging the product's market competitiveness.

An archetypal example of this intertwined effect of quality is Toyota and the Toyota Production System (TPS). This system, rooted in a set of practices and principals that enforce a culture of correction and learning, propelled Toyota past dominant American auto makers in both sales and financial performance. Starting at only 2% of US market share in 1970, Toyota crossed 13% of market share in 2006 [10] and is practically tied for first with 15% market share as of 2021 [42]. Toyota's profit also dominates its domestic competitors, averaging more than four-times higher earnings per vehicle than General Motors [43]. Quality of Toyota's product has not gone unrecognized by consumers, with a majority responding that Japanese cars were of higher quality than domestically produced ones [40]. Toyota's success is ascribed in part to the benefits it reaps from TPS, which counts outputs as being "defect free (ha[ving] the features and performance the customer expects)" as a central ideal [38].

Structured systems that seek to improve quality, often inspired by the Toyota Production System, are now found in many manufacturing companies. These systems

incorporate not only the hallmark learning behaviors and tools of TPS (such as "andon" boards and "kanban" inventory cards), but also fold in other frameworks: the Boeing Production System (BPS) purports to include elements of lean manufacturing, six sigma, and supplier relations [9]. Foundational to any of these flavors of quality management is data - so says the business maxim, what cannot be measured cannot be managed. In the case of Toyota and analyzing the performance of a worker bolting a seat into a new car, measurement is fairly tractable: a benchmark job time is set, and pulls of the andon cord indicate instances when that benchmark time is not met (with the time of the pull giving the amount of excursion from the expected time).

This type of structured data, whether it be numerical (as in timing a job) or categorical (as in counting the different reasons why a job could not be completed within the allowed time), is often key to process improvement. Difficulty arises when data is not in these easily manipulable forms but instead comes in formats like free-text prose.

Take, for example, a company's customer satisfaction survey. The company is interested in how their customers perceive the service they received, which one would often communicate in speech or in writing. Instead of accepting these formats, most surveys ask a customer to rate them on a numerical scale ("1 being the worst, 10 being the best") or by choosing a categorical value ("unacceptable," "average," "superior"). This is because it is far easier for the company processing these surveys to use structured values in determining its "customer satisfaction" metric; the alternative often requires significant human effort to read each response and assign such a numerical or categorical value.

This example is illustrative of a yet-evolving problem in many applications: how does one use unstructured data to the same degrees as structured data without expending inordinate resources?

1.2 Reactive & Proactive Processes

Many work processes can be broadly categorized as "reactive" or "proactive." Reactive processes are in response to some event and are characterized as being "after the fact" - for example: replacing a pump's bearings after they have worn out, or hiring for a position only after the incumbent has departed. Proactive processes act in anticipation of an event, often trying to either ensure or dissuade its occurrence: replacing a pump's bearings near the end of its life (but while still functional), or building a talent pool in anticipation of new work and open positions. With the proliferation of machine learning and Industry 4.0 thinking, "predictive" and "perscriptive" have been added to the lexicon: the anticipation of an event occurring based on the state of the system, and the identification of reactions that would impact an event's probability of occurrence, respectively [44]. While interesting, opportunities to improve reactive processes frequently exist even before adding these extra layers of sophistication.

Moving processes from reactive to proactive is often not trivial, as proactivity requires extra structure and diligence to do something consistently, and usually comes at some cost (either monetary or in opportunity). The value gained, though, can be large: in the case of the pump's bearing, two hours of planned downtime every other year may be far cheaper than a week's worth of lost production while waiting to procure an entire new motor.

Differentiating reactive from proactive can be nuanced, depending on the definition taken for each type of process. The International Civil Aviation Authority (ICAO), when describing hazard identification methodologies, chooses these: [27]

- **"Reactive"**: Concerned with "analysis of past outcomes or events" which are "indicative of system deficiencies."

- **"Proactive"**: "Involv[ing] collecting safety data of lower consequence events" and "analysing [...] frequency of occurrence to determine if a hazard could lead to an accident or incident."

Such definitions imply that proactive processes should (or must) attempt to manage a given event by managing the occurrence of its contributing factors; acting on past occurrences of the event itself is seen as being reactive. We do not find this distinction compelling, as it focuses on causality rather than intention. In this thesis, we choose "proactive" to be characterized by responses that are organized and informed by past occurrences, with the explicit goal of affecting (typically, reducing) further occurrences. In contrast, we consider "reactive" processes as characterized by single efforts with little to no memory of past occurrences, and with no consistent strategy to affect future occurrences.

As such, we take "proactive quality practices" to mean "intentional, structured, and repeated efforts undertaken to improve a data-informed measure of quality."

1.3 Boeing Commercial Airplanes Digital Quality Initiative

Boeing Commercial Airplanes (BCA) includes expanding data-driven analytics and building their enabling technologies as a current strategic focuses for its quality function. One area of interest within this focus is finding ways to better utilize data that is represented in less-standardized, more free-text forms. In some phases of the commercial airplane value stream (such as in the manufacturing phase) data is generally in a structured format of codes and attributes, with allowable values and relationships enforced by harmonizing business systems such as a manufacturing execution system (MES); metrics and analytics in these phases are well-studied and mature. In other phases (such as the in-service phase), data is commonly less-structured prose: examples include typed reports, e-mail message chains, and maintenance logbook entries originating from airline customers around the world. Relating these types of free-text data and extracting informational value traditionally relies on human effort to read and comprehend each example, and is a largely untapped source of analytical value relative to its structured counterparts. The volume of unstructured data that

Boeing both generates internally and collects from its customers each year makes a purely manual approach to information extraction or relationship-finding within these data sets unmanageable at any larger scale.

1.4 Extracting Information from Free-Text Data

The ability to automatically obtain information from free-text data is desirable in many business and operations applications, either for its analytical value or for other purposes. Certain applications may wish to extract key fields or features from written word, creating a more structured representation of the original data without requiring human effort to read and parse individual documents. Other applications consider contents more holistically, using models to transform documents into representations that, while not keeping any of the actual words or phrases used, maintain a sense of semantic similarity between documents.

In the aviation industry, extraction of "part type" and "condition" from maintenance logs is an example of creating a structured representation - typed records are processed by models that are able to both extract the part name (such as "nose landing gear") and the part's condition (such as "worn to limit"), which can then be used in ways similar to natively structured data (for example, creating categorical heatmaps of frequent maintenance findings) [16][26]. In public safety, emergency call centers have found uses for models that classify calls by "severity" and "priority," using location and call type information extracted from phone transcriptions as features [41]; research in the medical industry uses extractive methods to mine indications of adverse events from patient notes [14], as well as glean medication dosing amounts from clinician's instructions for pattern analysis [30].

Other methods do not rely on an explicit extraction step, but rather employ a transformation of the text as a whole. *Seale et al.* demonstrates an approach that directly transforms free-text rotorcraft maintenance records into a representation that maintains semantic relationships. These new representations are used as features for classification models which categorize maintenance records into health-based groups.

Results here show not only show high accuracy relative to manual "subject matter expert" scoring, but expand grouping coverage from only 10% of documents (the amount analysts are able to review manually) to 100% of documents [35].

1.5 Project Approach

The objective of this project is the exploration of areas where business processes that traditionally act reactively, due in part or in whole to the types of information available to them, can be moved to proactive through effective use of unstructured data. Processes that may benefit from broader or more consistent access to information stored in these formats are identified, and improvements that can utilize a combination of natural language processing (NLP) and machine learning methods to obtain and employ such information are trialed. In addition, we discuss the organizational considerations that may influence the success of these proposed changes.

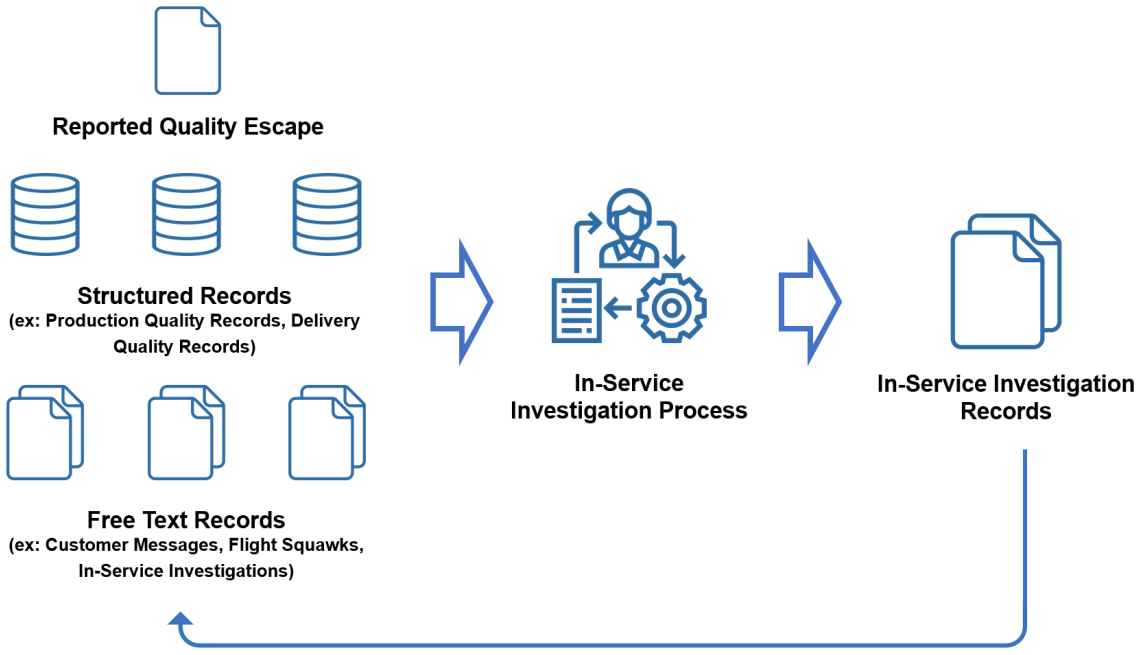

Figure 1-1: In-Service Investigation Cycle

Boeing Commercial Airplanes' "in-service investigation" (ISI) process is the concrete example this project explores: this process both utilizes many free-text documents in investigation execution as well as produces a largely free-text report as a final product

(Figure 1-1). Methods that help investigators collate and relate free-text documents reduce data discovery times, increase the consistency of data examination, and yield higher-quality investigations. The ability to identify trends in investigations (for example, what is the most common type of quality escape on electrical systems) links this data with similar quality trend analyses appearing earlier in the value stream, building a more complete picture of Boeing's total quality. We discuss the architecture and results of candidate frameworks that can offer these improvements.

1.6 Thesis Organization

This thesis is organized into seven chapters which further expand on Boeing and its linkage with quality practices, the particular algorithmic pipelines trialed and their results, and culminates with a discussion of overall effectiveness of the project's goals.

Chapter 2 discusses Boeing Commercial Airplanes in more detail, covering the commercial aircraft value stream, the historical and contemporary drivers of quality, and an elaboration on how quality practices differ within the value stream.

Chapter 3 considers the current uses of structured and unstructured data in quality programs, discusses the gap in utilization between these formats in proactive quality practices, and identifies several methods that may provide remedy. We formalize this discussion as a problem statement and describe an example use case involving Boeing's in-service investigation process.

Chapter 4 considers the technical aspects of this use case, exploring how data is represented in these investigations, which natural language processing and machine learning techniques will help extract the desired information, and what data cleaning and post-processing methods are necessary to enable such extraction.

Chapter 5 focuses on how to enable analysis and trending of quality topics seen in investigations. Two methods are explored: an unsupervised clustering approach, and a supervised label classification. We show some ability for unsupervised methods to group investigations by quality topic, but better performance from supervised classification methods.

Chapter 6 focuses on a way to improve the investigation process through use of a "helper" tool. We discuss the data mining process each investigation undertakes and how these searches seek results on the basis of similarity. We describe a tool to aid in these searches through automating data collection and similarity comparisons, and present an example of its use.

Chapter 7 discusses the findings of this effort in summary, including a review of the in-service investigation process as an appropriate use case, recommendations for how such methods can succeed within Boeing, and concludes with a perspective on broader sentiment towards proactive quality and machine learning as an enabler.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Boeing Commercial Airplanes & Quality

Boeing Commercial Airplanes (BCA) a major supplier of commercial passenger and freight airplanes to the global market. As a manufacturing company that has produced product for over a century, the concept of quality and quality management is well established. The matrix organization within BCA, which separates roles by value stream phase and airplane program, can complicate efforts to pervasively share quality information. As a company who uses product quality as a market differentiator, BCA has a strong business interest in any improvements to quality practices.

2.1 Boeing Commercial Airplanes

The Boeing Company, founded as Pacific Aero Products Corp, traces its origins building aircraft in the Puget Sound, WA area to the 1916 launch of the Boeing Model 1 seaplane [12]. Modern-day Boeing Commercial Airplanes (BCA) extends this legacy with aircraft that have long-held synonymous associations with commercial aviation and American engineering: its 737 narrow-body airplane program holds the world record for "Most Produced Commercial Jet Aircraft Model" [1] on a design that was first introduced in 1967; the "Queen of the Skies" 747 wide-body airplane held the

[1] At the time, a record of 10,000 airplanes produced [11].

title of "Largest Passenger Plane" for almost four decades and has seen life as a Space Shuttle carrier [18], airborne observatory [29], and as US Presidential transport "Air Force One."[2]

2.2 Commercial Airplanes Value Stream

The commercial airplanes value stream (Figure 2-1) can be broken into three main phases: production, delivery, and in-service.

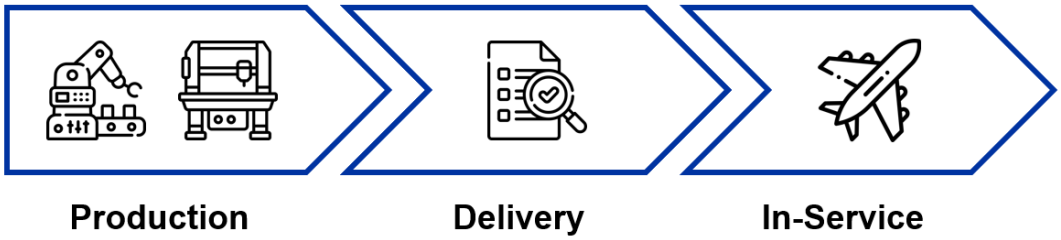

Figure 2-1: Commercial Airplane Value Stream

Production, also known as manufacturing, comprises the tasks of building and outfitting a new aircraft. Major structures of the aircraft (such as the fuselage barrel sections, wings, and stabilizers) are built on-site or acquired from a supplier and assembled in the initial portion of the build; latter portions involve adding electrical and hydraulic systems, interiors, and exterior painting. Certain functional testings (such as pressurization testing) are carried out during the production process. Work processes in production are dictated by "installation plans" (IPs): these documents list the ordered steps, materials, and checks required in each build process. Boeing offers high levels of customization to its airline customers, making the set of installation plans used for each aircraft build dependent on its customer and ultimate configuration. Maintaining this relationship of build planning, along with associated records of test results and quality items, are a critical feature of the manufacturing execution system (MES). Time to complete the production process for an aircraft is dependent on complexity, customer demand, and manufacturing capacity. The smaller, high-demand

[2]The next generation of "Air Force One" aircraft are being manufactured, based on the 747-8 model [12].

737 program produces 26 aircraft per month; the larger, lower-demand 747 program produces a single aircraft every other month.[3]

Formal check-out and certification of a new aircraft comes in the delivery phase. The aircraft at this point is fully assembled and ready to be tested in anticipation of obtaining an airworthiness certificate, officially becoming a commercial aircraft. Delivery checks include initial fueling, engine runs, and avionics testing and is punctuated by two key milestones: the B-1 (Boeing 1) and C-1 (Customer 1) flights. These are the first times Boeing and the aircraft customer, respectively, fly the plane; issues found during flight (associated or unassociated with any specific test) are recorded as "flight squawks," which are resolved prior to aircraft turnover. After all outstanding actions are completed and an airworthiness certification is obtained, the process culminates with the actual delivery of the aircraft to its customer for entry into revenue service. Delivery operations run in parallel, with several aircraft in various phases of the process at any given time; delivery process times are typically in the range of several weeks [12].

Post-delivery, the remaining life of an aircraft is referred to as "in-service." While Boeing no longer has control of or regular interaction with the aircraft, it does maintain customer support in a manner similar to many other equipment manufacturers. Support can range from providing maintenance instructions for scenarios not covered in existing manuals to responding to quality discrepancies found during service that the customer believes were introduced by Boeing during manufacturing. This support is a mixture of contractual obligations and optional after-market services.

These value stream phases are largely concerned with the building and supporting of individual aircraft. In addition, BCA also hosts large airplane design engineering and product development functions. These groups work to define future generations of aircraft, either as minor-model types of existing airplane programs (such as the 737-MAX9 and 737-MAX10), or as new major models (such as the 787). While feedback loops between design engineering and the production system do exist, resolutions to quality issues requiring an aircraft design change are in the minority.

[3]Per Boeing Commercial Airplanes Fact Sheet as of December 31, 2021 [12].

2.3 Commercial Airplane Quality

Within the commercial airplane marketplace, Boeing positions itself as the supplier of premium product offerings.[4] The quality of these products is thus tied to reputational and financial metrics: products that do not live up to the quality commensurate with their price tag may cause airlines to move to a competing product, and poor quality manifesting in rework only subtracts from the margin Boeing makes on every aircraft. Recent events have drawn the topic of commercial airplane quality, specifically the quality of Boeing aircraft, to the forefront in the eyes of both government agencies and of the traveling public.

Regulators have found concerns within Boeing's manufacturing environment on several occasions. In May 2021, Boeing was fined $17 million for quality escapes in the production of the 737 related to part quality and certification [13]. Manufacturing quality concerns on the 787 program caused Boeing to halt deliveries of their newest airplane program in May 2021 [37], after being fined $6.6 million in February 2021 for gaps in compliance and quality systems that included the 787 program [21].

Two crashes of the 737 MAX aircraft within 6 months, Lion Air 610 and Ethiopian Airlines 302, killed 346 people and caused the FAA to ground the MAX derivatives in March 2019. Public willingness to travel on a MAX aircraft noticeably waned even before this intervention - on the day of FAA grounding, travel website KAYAK released a feature that allowed its customers to filter out flights on particular aircraft models [12].

2.4 Practicalities of Quality Information

Boeing Commercial Airplanes is a matrix organization, differentiating teams and responsibilities by strong axes of "value stream phase" and "airplane program." With respect to quality, value stream phases share common methods and tools for identifying

[4]The Boeing 737-MAX8 and Airbus A320neo occupy a similar portion of the product landscape in terms of range vs passenger capacity. The average price of a 737-MAX8 is $121.6 million[12]; the average price of an A320neo is $110.6 million [8].

problems and recording their resolutions, but information and its associated knowledge does not necessarily pass to adjacent airplane programs. For example, all airplane programs have quality inspectors who verify correct execution of manufacturing steps, but these inspectors are assigned to a specific program (Figure 2-2). Dominant sources of quality information and the methods used to record observations and reactions differ in structure based on the value stream phase.

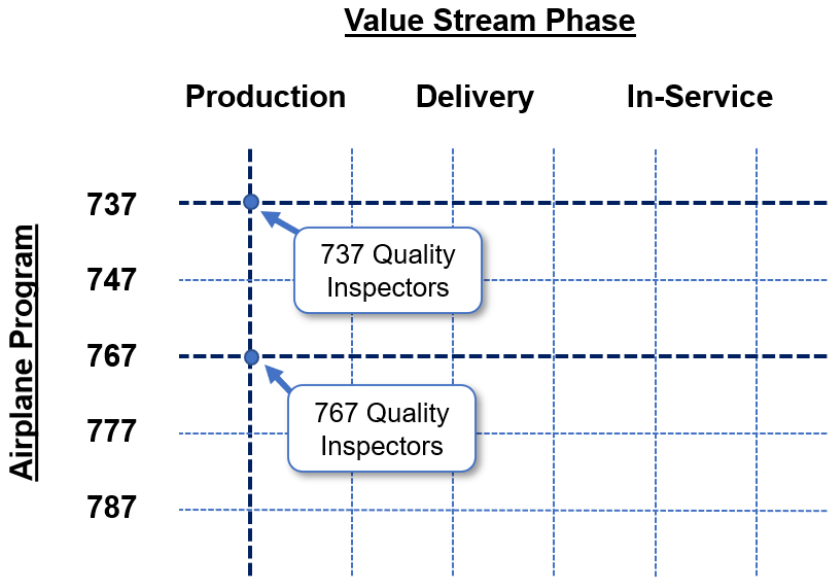

Figure 2-2: BCA Matrix Organization Alignment of Equivalent Jobs

In the production phase, quality information is largely records of defects in the build process as discovered by quality inspections. As this phase is guided by installation plans, these errors become associated with the IP that introduced or found the error; this type of information (such as "fastener - rivet; loose") is typically recorded through the use of standardized codes. Quality information in the production phase, due to its highly structured nature, is the easiest to aggregate: filtering by IP or aircraft will yield a list of quality issues identified, which can be further grouped by features such as flow day, system type affected, or error cause.

The delivery phase starts to introduce free-text as a more common method of recording quality information. While certain portions of the delivery phase are guided by test plans, there are opportunities for pilots to note anything they find unusual or incorrect in the form of a "flight squawk." These are narrative descriptions of the

concern and may not be associated with any particular test or scenario (for example: "rattling heard behind lavatory wall during cruise"). The corrective action may be documented in free text as well, with potentially little information captured in a form that is immediately relatable to other quality findings. The rate of findings in this phase is lower than the rate in production but are not rarities, as quality inspectors are still purposefully looking for escapes.

Quality information in the in-service phase is almost entirely free-text narratives. Aircraft are now out of the hands of Boeing, so notification of quality concerns come directly from the airline customer in a format similar to emails. Instances of quality issues that may have been introduced during the production and delivery phases are formally reviewed through the in-service investigation (ISI) process (discussed more in Section 3.5). The result and resolution of these investigations are also stored as free-text narratives, with some key information (such as part numbers and major system affected) recorded as metadata. The free-text nature and relative sparsity of these reports makes simple aggregation challenging.

Despite differences in the methods of obtaining and recording quality information in the different phases of the commercial airplane value stream, there is commonality in the system of record used to store these documents. The manufacturing execution system serves this purpose. This platform contains the documentation associated with each aircraft build and its associated quality records, from installation plans to ISI reports. While the structure and features are common for each phase, the usage and emphasis can differ - for example, entering metadata describing how a discrepancy was introduced is emphasized in production records, while it is more often missing or non-descriptive in ISI records.

The differences along the axis of airplane program are less severe relative to differences between value stream phase. As quality practitioners in each airplane program use the same tools to obtain the same outcomes as their counterparts within a value stream phase, the methods look quite similar. The barriers seen with this axis of the matrix structure are related more to the flow of information. Inter-program collaboration between sibling teams is often low, as airplane programs have low overlap

in the way of detailed manufacturing. This leads to an impression that quality trends are not related between programs (ie, a problem area for 737 is because of that build and those resources, and is not related to problem areas on 777).

This combination of different quality information schemes laterally across value stream phases and low organizational drive to compare vertically across airplane programs creates a challenging environment for views of total, holistic quality.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Considerations & Problem Statement

Quality monitoring and improvement practices create data in both structured and unstructured forms. Data in structured forms are more often what these improvement processes, especially proactive ones, are based on despite the wealth of informational value contained in unstructured data. Merging information from these types of data can both improve existing proactive methods and be the catalyst needed to change traditionally reactive processes into proactive ones. We take the in-service investigation process, which both consumes and produces largely free-text data within Boeing, as a candidate for transformation and consider several natural language processing and machine learning methods to assist in extracting information from these sources.

3.1 Structured Data in Quality

Data-driven quality management programs seek to monitor performance of a system, using deviations from a "norm" as indicators of potential quality issues. Conventionally, measurements that inform these types of surveillance are rooted in structured, often numeric, forms of data. Statistical process control (SPC), for example, is a common framework for quality control which leans heavily on numerical summaries of mean and standard deviation to draw insights.

Quality processes in Boeing's production phase rely on these types of structured records, which are generated in large quantities for each aircraft built. For example,

production staff may wish to understand what quality performance is like on a certain flow day. [1] Records from installation plans that occur on that flow day could be selected from the set of available quality documents, then further aggregated by an intermediate field such as "type" (e.g., structures, electrical). The resulting aggregation is informative both temporally and distributionally: counts of defects found per time span (or per aircraft build) shows how quality is improving or declining, and the associated metadata helps highlight problem areas. For example, finding many quality records that mention different types of loose mechanical fasteners (which include rivets, bolts, and screws) points to a different intervention than would finding many records which mention only loose rivets.

The combination of record volume, specificity, and completeness allows even simple aggregations of structured quality documents to produce insightful information. Boeing frequently steps beyond these methods, employing more sophisticated statistical and scientific methods (such as SPC) on its data to understand quality performance.

3.2 Unstructured Data in Quality

Another representation of quality data is found in less-structured forms, where information is captured in written word. Examples of processes in this category, such as maintenance log entries or short "trouble" tickets (like flight squawks), may be rooted in practices that did not anticipate aggregation when they began and that may have significant organizational or technical barriers impeding movement to a more structured format (for example, expecting global airlines of varying sizes and sophistication to move to an electronic maintenance system). Other instances involve records that simply are best represented as free-text. Customer messages describing the circumstance and presentation of a quality issue that has been found are better described in prose than with pull-down menu choices. Though it can be beneficial to move quality reporting towards more structured formats that better support native aggregation, a question of the historic data still remains - information from records

[1] "Flow day" refers to a particular day in the airplane build.

before the "structured" switch may remain underutilized unless effort is expended to re-cast them in the new, structured format.

Within Boeing, quality records from later phases of the value stream are substantially less structured than those from earlier phases. These documents largely come in the form of free-text narratives, where additional information may be simply appended to an ever-growing block of text. Drawing insights with these types of records is observed to largely comes from two means: metadata and "insider knowledge." Many records, though consisting primarily of free-text, are labeled with attributes that attempt to extract some summary or relational values that may be useful for aggregation. In the example of in-service investigations, each investigation is labeled with metadata that includes the general system involved ("electrical," "interiors," "lights"), the age of the aircraft, and the datum part number. An alternative to formal aggregation is "insider knowledge" - with ISI's, investigators gain an awareness of what types of issues are reported more often and have an internal affinity for the "hot areas." This type of analysis, however, is entirely dependent on individual investigators having both the awareness of the contents of many investigations and the wherewithal to draw connections and insights, and is hardly ever formally documented in depth.

3.3 Extracting & Utilizing Information from Free-Text Data

Advancements in the field of natural language processing (NLP) and the prevalence of NLP tools and pre-trained models continuously improve the accessibility of such techniques to non-expert users. Freely-available libraries such as "HuggingFace," "Natural Language Toolkit" (NLTK), and "Gensim" open a variety of NLP methods to those familiar with the Python programming language. While NLP greatly assists with the extraction of information from free-text data, there is no one singular approach - different NLP methods produce different types of results. Common examples are token extraction (for example, extracting part names and conditions from aircraft

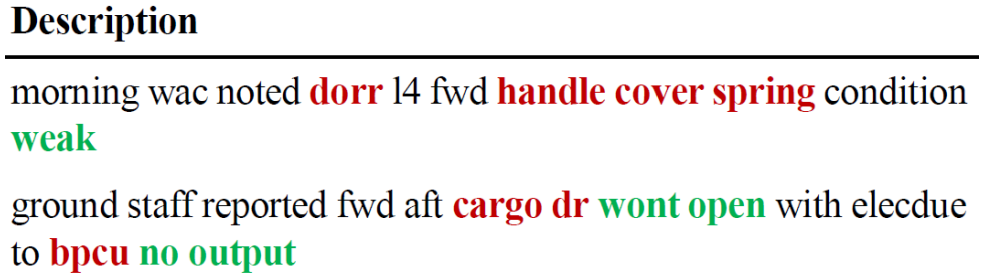

Figure 3-1: Extracting Part (red) and Condition (green) from Written Maintenance Logs (source: *Niraula et al.* [26])

maintenance records), summarization (creating shorter text summaries from a larger document), and sentiment analysis (classifying what sentiment a piece of text is conveying, such as "appreciation" or "aggrevation" in a customer satisfaction survey).

The type of technique used to extract information from free-text is dependent both on the way information is encoded in the data and what the end use of such extracted data will be. Previously mentioned is the example of using token extraction to obtain part names and conditions from written maintenance logs (Figure 3-1). The format of this type of data, while still in prose, has an expected structure: sentences include a type of part and a description about the reason maintenance was performed on that part. Obtaining these tokens also has an expected use: each record can now be represented as a set of data (namely, the part and the maintenance reason), which can be utilized in quality processes suited for structured, categorical information.

Point uses of such extracted data are interesting on their own, but greater value can be derived when such information is incorporated into broader programs that may traditionally be informed only by structured data. For the maintenance log example, one candidate within Boeing is its "aircraft health management" (AHM) product. With AHM, properly equipped aircraft record a variety of performance values, error codes, and diagnostic data which are then streamed in real-time from the aircraft [24]. This product provides operators with a proactive way to manage their fleet's health - new issues (unexpected diagnostics messages, measurements deviating from their expected values) can be readily identified and resolved before they impact an aircraft's ability to operate. Including maintenance log information, historically

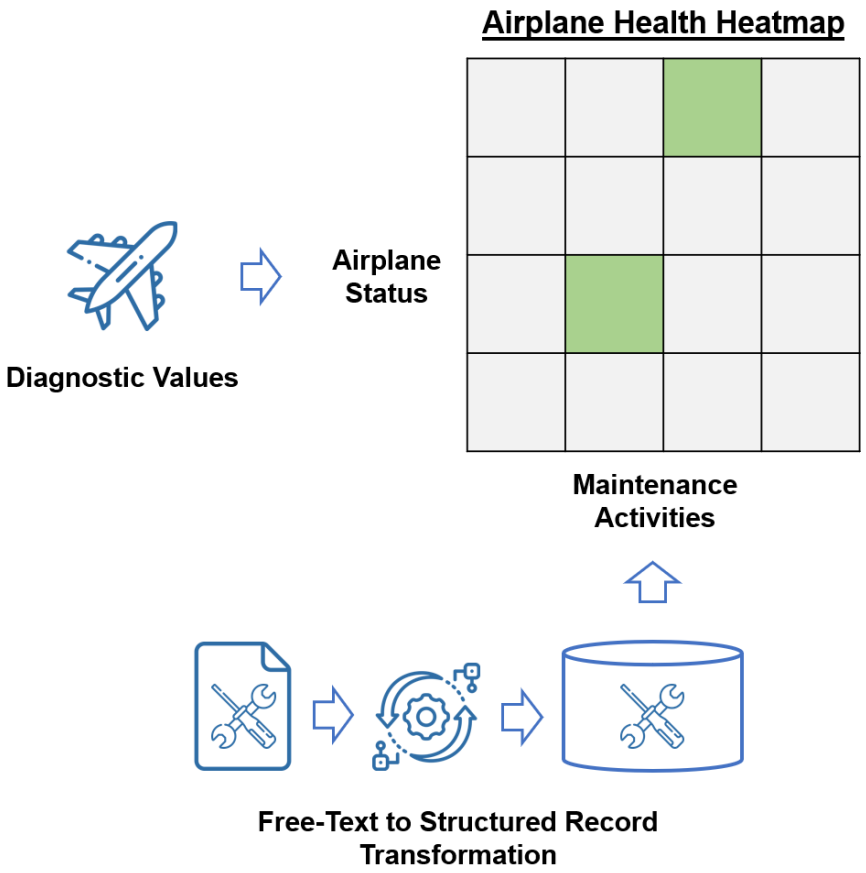

Figure 3-2: Potential Heatmap from Structured Diagnostic Data and Unstructured Maintenance Data

analyzed separately, with these types of "digital maintenance" data builds a more complete representation of a plane's health. Relationships that could not be found easily when structured and unstructured data were separated, such as finding that a particular diagnostic message consistently precedes a maintenance activity to replace a certain part, are now much more readily obtainable (Figure 3-2). This new mix of information from structured and unstructured data improves the coverage and effectiveness of an existing proactive process.

3.4 Problem Statement

Quality management practices involving structured data, including those at Boeing, are often mature and enable proactive efforts to improve quality performance. Gaps

appear when data reflecting quality is not in these structured forms and cannot be easily aggregated or summarized; such data tends to not inform proactive quality processes but instead is acted on more reactively. Effectively exploiting this unstructured information, which often contains large amounts of free-text, can improve the results of proactive quality management practices.

The process we use to demonstrate this theory is in-service investigations. This process investigates potential quality escapes found and reported by Boeing's airline customers, and has several hallmarks of a suitable candidate:

- **Free-Text Data**: The ISI process both consumes and produces large amounts of free-text data - relating information in these records and finding similarities is left to human effort and judgment.
- **Reactive**: The ISI process reacts to a customer's request, with action focusing on the limited issues presented - effort follows open investigations and dwindles when investigations conclude.
- **Disconnected**: Summarizations of ISI results (both contributing issues and resolutions) are not consistently constructed nor related to summarizations from other parts of the value stream, despite the similarity of their insights.
- **Reinforcing**: Part of the investigation process involves reviewing previously completed investigations to look for similar instances - improving ways of comparing investigations and finding more similar examples improves the process overall.

We choose two aspects of the ISI process to explore the benefits of improved unstructured data use: how to better analyze subject quality topic trends, and how to better find similarities in completed investigations.

Quality Topic Trends

Each in-service investigation is based on something an airline customer has found and believes to be a quality escape from Boeing. The broader class of this finding can

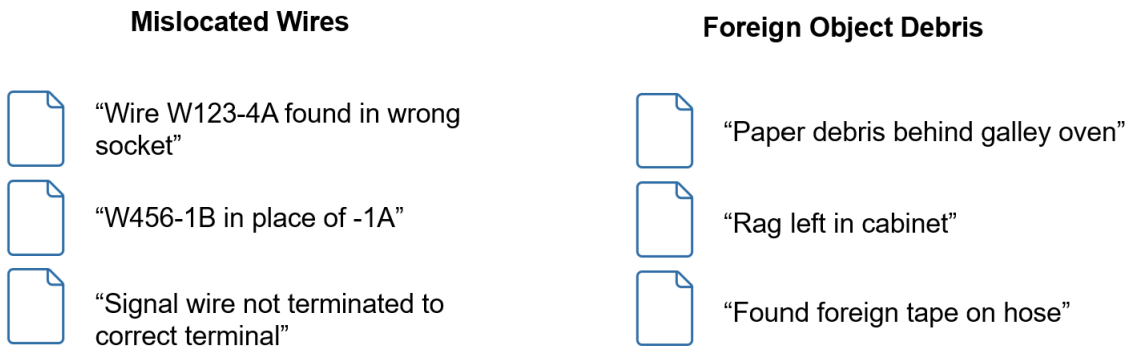

Figure 3-3: Example Quality Issues Grouped by Quality Topic ("Mislocated Wires," "Foreign Object Debris")

be referred to as a quality "topic" (as in, the topic of the investigation), as illustrated in Figure 3-3. Quality topics are more informative for grouping than exact quality escapes, as we find these rarely repeat. Trending these quality topics provides valuable information: consistently higher occurrences of a topic may indicate an area that needs more attention regarding quality improvement, while a rising occurrence rate of a topic may indicate degradation in existing quality management systems (Figure 3-4).

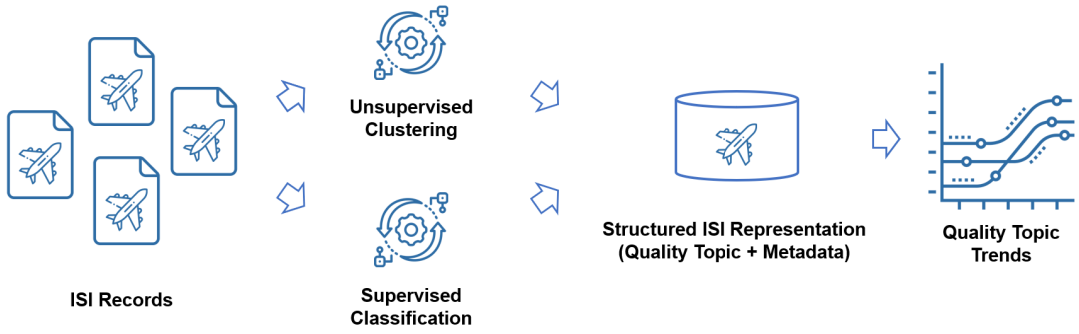

Figure 3-4: Workflow to Analyze In-Service Investigation Trends through Quality Topic Grouping

Two approaches are considered for obtaining this type of relationship:

- Unsupervised Clustering: In-service investigations are represented by an NLP embedding of their contents and are segmented by their mutual similarities into groups of quality topics.

- Supervised Classification: In-service investigations are classified as one of a list of pre-defined quality topics based on their NLP representations.

Finding Similarities in Completed Investigations

The ISI process includes a review of previous investigations to understand if the quality topic seen in a new investigation has been reported before. Keyword searches or metadata filtering of potentially related investigations are the main means of performing these searches, but results depend on the prowess of the investigator to identify correct search terms and appropriate filters. The ability to help investigators both arrange relevant ISI data and provide an NLP-powered indication of similarity through a purpose-built tool improves the consistency and completeness of data discovery.

3.5 In-Service Investigations

The in-service investigation process is part of the mechanism Boeing's customers use to report quality concerns found on their delivered aircraft. Investigations are initiated by a customer message detailing a problem they believe to be caused by an escape of Boeing's quality program: examples include tools left on the aircraft, prematurely peeling paint, and nicked wires. Upon receipt of this message, an investigator will review the datum aircraft's factory and delivery history, examine other investigations for instances of similar issues, and ascertain the containment and remediation actions necessary to prevent a similar quality escape from reoccurring (for example, re-writing an installation plan, employee coaching, or supplier engagement). The intermediate findings and conclusion of each investigation is captured as a written narrative; this is recorded, along with a set of standardized finding and corrective actions codes, in the MES (Figure 3-5).

Drawing deeper insights from in-service investigations poses several challenges. One is the unstructured nature of investigation reports: key information on what the quality problem is and how it was introduced is stored in the free-text narrative.

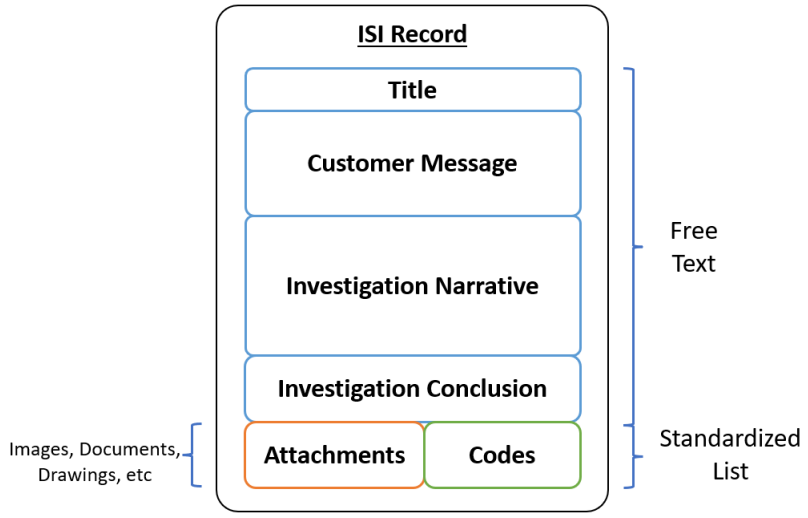

Figure 3-5: In-Service Investigation (ISI) Record Structure

Standardized codes do expose some of this information in a more exploitable manner, but they suffer from being incomplete (one or more codes not entered) or overly-broad (using options like "part" and "incorrect" rather than more descriptive codes such as "wire bundle" and "mislocated"). Grouping investigations by the specific part involved is similarly inadequate as exact part numbers rarely repeat (Figure 3-6). Confining the definition of "repeat occurrence" to exact matches of part number and location is overly narrow, especially to external parties such as customers and regulators: answering "how often is this particular bolt found under-torqued" does little to convincingly show Boeing's history and improvement of all loose fasteners.

Organizational challenges to broader analysis involve the airplane program-centric view of in-service investigations. Investigators and quality teams typically focus on only one airplane program, building knowledge around what issues often occur for that type of aircraft; comparing those problems to other programs generally is the exception rather than the rule. Given the number of planes that Boeing delivers and the complexity of a modern commercial jet, in-service investigations are quite infrequent (Figure 3-7): an investigation (which could be for something as minor as a scratched armrest) is requested, on average, for only one in every 40 airplanes delivered. Segmenting this information by program boundary reduces visibility into

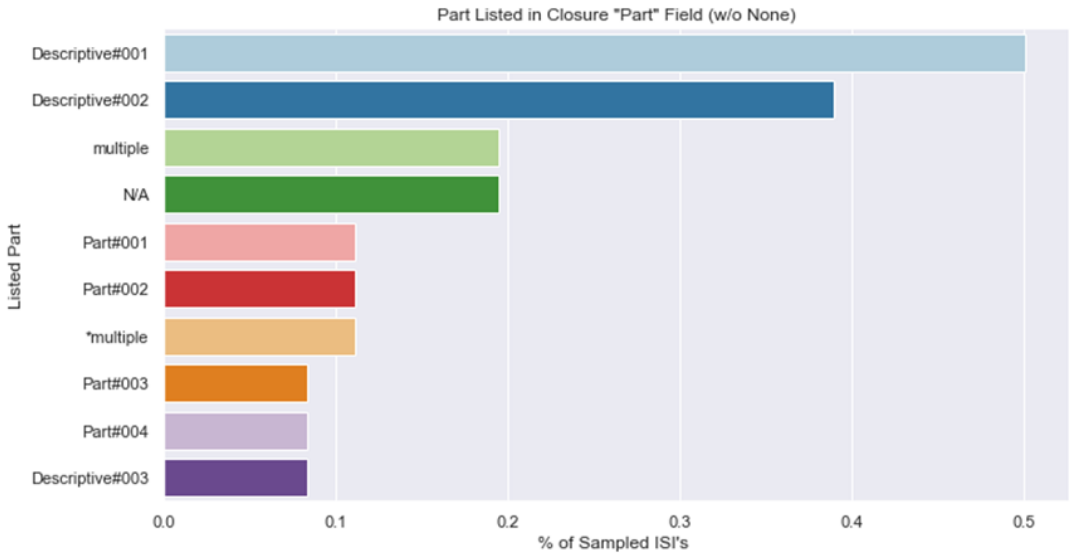

Figure 3-6: ISI Repeat Part Number Distribution ("Descriptive" refers to a categorical part, ex: "paint")

an already small data set; systemic issues may be missed because their program-level components seem relatively small. This is especially problematic for quality topics that do have common factors between airplane programs. Electrical wire bundles, for example, for a majority of airplane programs are produced in a single facility - considering quality topics regarding wiring issues on an airplane program-only level ignores this relationship.

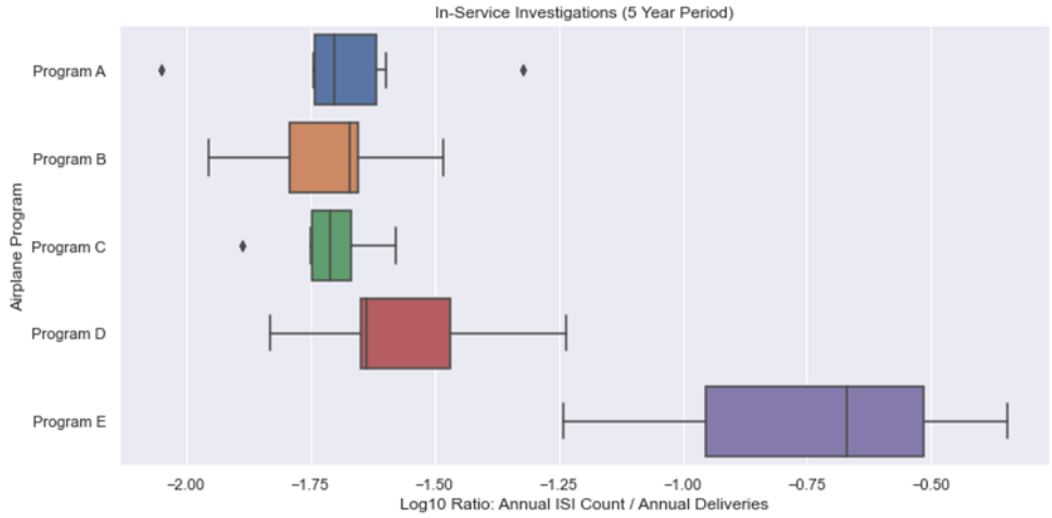

Figure 3-7: In-Service Investigations - Annual Count of Investigations Opened per Airplane Delivered

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Data, Technical Considerations, and Data Transformers

The specific techniques used to extract information from unstructured data depend on the form such data takes, as well as the desired final representation. With in-service investigations, data consists of multi-word phrases or sentences, where semantic relationships are important for capturing information as different phrases are used to describe similar ideas. Sentence "embedding" techniques perform well in this circumstance, but can lead to high-dimensional results. Choosing the right combination of embedding technique and dimensionality reduction method is necessary to produce a usable representation of unstructured, free-text data.

4.1 In-Service Investigation Data Set

The data set explored in this thesis is a collection of completed in-service investigations. Each record in this set represents a single investigation, consisting of close to 100 separate data fields detailing the state and conditions of the investigation, the history and configuration of the datum aircraft, and free-text and categorical summarizations of investigation resolution. Key data used for this study is listed in Table 4.1. This data set includes approximately 6,100 records over six airplane programs.

Though most records have values for each of the key data fields, examples are found

Feature	Description
Title	Descriptive title for the investigation, written by the investigator
Reported Date	Date that the investigation was opened
Aircraft Information	Model, Serial Number, Flight Hours, Flight Cycles of the investigation datum aircraft
Location/System information	Standardized ATA Code and descriptive Location code of area quality issue was found
Part Number	Specific part involved (if any)
"Found" and "Produced" codes	Standardized codes to describe the condition the quality concern was found in (such as "fastener, loose") and the means the issue was introduced to the aircraft (such as "hole, not per specification")
Initiating Customer Message	Free-text message received by Boeing that started the investigation

Table 4.1: Key ISI Document Features

with missing data. As such, a subset of data is often used here for model training and evaluation (such as selecting only particular "found" codes). While additional manual effort is spent attempting to find all relevant samples of these subsets, it is possible that examples are missed.

4.2 Transforming Free-Text

A requirement for utilizing information from in-service investigations is the ability to transform the free-text portions of each investigation into an appropriate representation, whether it be for clustering, classification, or direct similarity comparison. Within natural language processing, many potential transformation methods are available.

One class of method is based on how frequent a word occurs in a document or the relative frequency of different words occurring in a set of documents. A popular technique that follows this pattern is "term-frequency inverse document-frequency" (tf-idf), where documents are parsed into a set of tokens (consisting of one or several consecutive words) and an "informational value" statistic is calculated for each token [6]. This statistic increases as the token appears more times in a document but decreases

Figure 4-1: SBERT Embedding and Example Spectral Clustering (n=3) of Common ISI Corpus Words

as it appears in more documents, following the idea that tokens that appear only in some documents carry higher informational value. While a simple transformation, this type of technique contains no sense of semantic relationship - words or phrases that mean the same thing are still treated as separate tokens.

Another class of method seeks to preserve the semantic relationships between words or phrases in a document. These techniques transform text into a vector space representation, called an "embedding," that represent words as a multi-dimensional numerical vector. Words with similar semantic meanings occur in similar portions of the vector space; a numerical value for this similarity between two words can be obtained by computing the cosine similarity of the two embedding vectors. A number of vector representation frameworks and pre-trained models are readily available, such as "Transformers" [45] and the "Bidirectional Encoder Representations from Transformers" (BERT) [15] model, respectively.

A separate class of transformations involve classifying and extracting specific tokens from text. The maintenance log example discussed previously, where tokens representing "part" and "condition" are identified, falls in this category. These types of methods typically require greater attention and expertise to craft an appropriate model; given the resource constraints of this study, this class is not considered.

Each feature in an ISI record is a free-text word, group of words, or full sentence.

Some features, such as the investigation's title, are unconstrained and represent an investigator's impression of descriptive information. These types of features have no consistent syntactical structure ("nose landing gear door hinge found broken" vs. "broken hinge found common to nose landing gear door") and may use different words to convey the same meaning ("broken," "failed," "inoperative"). Other features, such as "found" and "produced" codes, take values from a list of pre-defined choices. These lists, however, have no relational structure (values exist solely as a possible choice) and contain potentially interchangeable values ("wire bundle," "electrical wire harness").

Given these types of features within the ISI data set, the most appropriate method is one that can transform multi-word pieces of text into a representation that maintains a measure of the original text's semantic meaning. The "Sentence-BERT" (SBERT) [32] transformer is chosen for this task - capable of quickly transforming sentences into a fixed 768-dimension embedding space, SBERT also has a number of pre-trained models available for immediate use. Throughout this thesis, SBERT and the "all-mpnet-base-v2" model [28] are chosen for free-text embedding operations, based on strong performance representing ISI-corpus words (Figure 4-1). Both tf-idf and similar word frequency transformations are used in some instances as a comparison to embedding representations.

4.3 Data Cleansing

Many features within the ISI data set are human-written and contain a mixture of abbreviations and non information-adding phrases that require correction before text transformation can be applied. Data cleansing is performed in two steps: acronym expansion, and phrase removal.

4.3.1 Acronym Expansion

Acronym expansion attempts to find acronyms and shorthand phrases commonly used in written text and replaces them with full phrases (Table 4.2). This step is more critical for models that use embedding transformations as features, as these methods

Abbreviation	Full Phrase
12MAQ	12 Month Airplane Quality
AOG	Aircraft On Ground
NLG	Nose Landing Gear
W/B	Wire Bundle
FSTNR	Fastener

Table 4.2: Sample of Common Abbreviations

rely on the new representations capturing the meaningful similarity between words and phrases. As the pre-trained model employed is not specifically trained on aviation industry corpus, any acronym encountered may be interpreted as an out-of-vocabulary word (yielding no transformation and no informational value) or, worse, inferred as having a different meaning and obtaining erroneous informational value. An example of this is the acronym "LOPA." The aviation industry uses this to stand for "layout of passenger accommodations" (i.e. a plane's interior configuration). The chemical industry, however, uses this to stand for "layer of protection analysis," which refers to a process for formal hazard identification.

Acronym expansion is applied to features used by word frequency transformations as well. Here, expansion is less critical as these methods do not consider any word meaning similarity, but does help create a more accurate transformation in instances where both the acronym and full phrase are used in a document. Research on more advanced methods to normalize and extract root phrases (particularly relating to part descriptions) is being conducted at Boeing [22], but is not employed in this trial.

4.3.2 Phrase Removal

Phrase removal attempts to eliminate words and phrases that are not expected to contain information that relates to quality topics. These phrases are often added by the customer or investigator to emphasize the impact of a quality issue, but do not help differentiate one quality topic from another. Phrase removal is typically only applied to the "title" feature in an attempt to create a more succinct description of the investigation. Dropped phrases include:

- "12 Month Airplane Quality" (12MAQ), referring to a quality regime that focuses on the first 12-months of an airplane's service life.

- "Aircraft On Ground" (AOG), referring to the aircraft being in a state where it is not available for revenue service.

4.4 Dimensionality Reduction

Dimensionality reduction is a transformation that changes data from a high dimensional space to a lower dimensional space, while still retaining a level of the variability seen in the original form. In NLP, reducing dimensionality is advantageous for improving performance (if not pure feasibility) of algorithms that use embeddings as features [31].

The ISI data set consists of 6,100 records, though most individual airplane programs contribute less than 1,000 of these records; the SBERT model, in contrast, produces 768-dimension embedding representations. Considering a subset of investigations consisting of one airplane program and only two features would yield more columns ($2 \cdot 768 = 1,536$) than rows, likely thwarting clustering attempts. Reducing the dimension of the embedding space to a reasonable degree is necessary for use of more than one or two features.

We analyze five classes of dimensionality reduction technique for this purpose:

- Principal Component Analysis (PCA)

- Spectral Embedding

- Kernel PCA (KPCA) (linear, cosine, and full-cosine kernels)

- Sparse PCA (SPCA)

- Uniform Manifold Approximation and Projection (UMAP) (linear and cosine kernels) [33]

4.4.1 Process

While each technique aims to maintain the variability of the original data, they do so through different means and yield different representations. In the case of text embeddings, we seek a representation that conserves the mutual similarities of the original feature space - otherwise, the power of the embedding transformation is lost.

We judge performance of each technique against this goal by first transforming a list of words and phrases into embedding space and computing mutual similarity between the list's members. This new embedding representation is then reduced by each technique to a target dimension (given as a hyperparameter), and mutual similarities are again computed. Comparing these two similarity matrices allows us to determine conservation of mutual similarity.

```
moves := 0
while array a != array b:
    i := argmin(a_i != b_i)
    u := b.drop(b = a[i]).insert(a[i], i)

    j := argmax(a_j != b_j)
    v := b.drop(b = a[j]).insert(a[j], j)

    if KendallTau(a, u) > KendallTau(a, v):
        moves := moves + (i - arg(b = a[i]))
        b := u
    else:
        moves := moves + (arg(b = a[j]) - j)
        b := v

return moves / length(a)
```

Figure 4-2: "Sort Order" Pseudo Code

Similarity between these matrices is computed using a novel implementation of edit distance, called "sort order" as summarized in Figure 4-2 (see Appendix A for full code). Whereas edit distance determines the difference between two sequences by the number of additions, deletions, and swaps needed to turn one into the other, sort order only considers swaps as the two sequences, by construction, are re-orderings of the

same elements. Sort order sums the number of element swaps needed to turn the first sequence (sequence a) into the second sequence (sequence b); swaps are weighted by how many positions the element is being moved to capture the "wrongness" of the first sequence. The final sum is divided by the length of the sequence as a normalization. This procedure is performed on each pair of rows in the similarity matrices and the cumulative sum is returned as the final metric. Lower values are better, as it indicates smaller changes are needed to sort b into the order of a.

Sort order computes in $O(n)$ time by assuming that the next best swap is either correcting the placement of the first misplaced element (creating sequence u) or last misplaced element (creating sequence v). Kendall's τ [34] coefficient is used to evaluate the sortedness of u and v compared to a; the option that improves the sortedness more becomes b, and the distance the element moves is added to the running total.

4.4.2 Results

This metric is applied to several sets of ISI feature corpora, including "Air Transport Association" (ATA) chapter, "description" codes, and "feature" codes (Table 4.3). A typical comparison result is shown in Figure 4-3. Each trend line represents a different dimensionality reduction technique: the number of components chosen for the comparison trial are given on the x-axis and the resulting metric value is given on the y-axis - a lower value is better, as it indicates fewer swaps are needed to change the transformed ordering back to the original ordering. Flat lines (such as KPCA full cosine and SPCA) are methods that do not take a number of components as a hyperparameter. KPCA cosine kernel and PCA repeatedly show best results for methods with a configurable number of components, with break-over performance observed for $20 \leq n_{components} \leq 30$. The ability to configure the number of components is desirable as it allows for control over the dimension of the final space.

Kernel PCA (cosine kernel) is chosen as the preferred dimensionality reduction technique, and is used typically with $n_{components} = 25$.

Corpus	Description	Examples	Corpus Length
Air Transport Association (ATA) Chapter	Standardized numbering method for aircraft systems	"Fuel," "Instruments," "Lights," "Wings"	50
"Feature" Code	Set of possible values used to describe the component involved in a quality discrepancy	"Actuator," "Bolt," "Cable," "Protective Coating"	222
"Description" Code	Set of possible values used to describe the state of a quality discrepancy	"Chipped," "Length Oversized," "Not per Drawing," "Torn"	238

Table 4.3: Dimensionality Reduction Evaluation Corpora

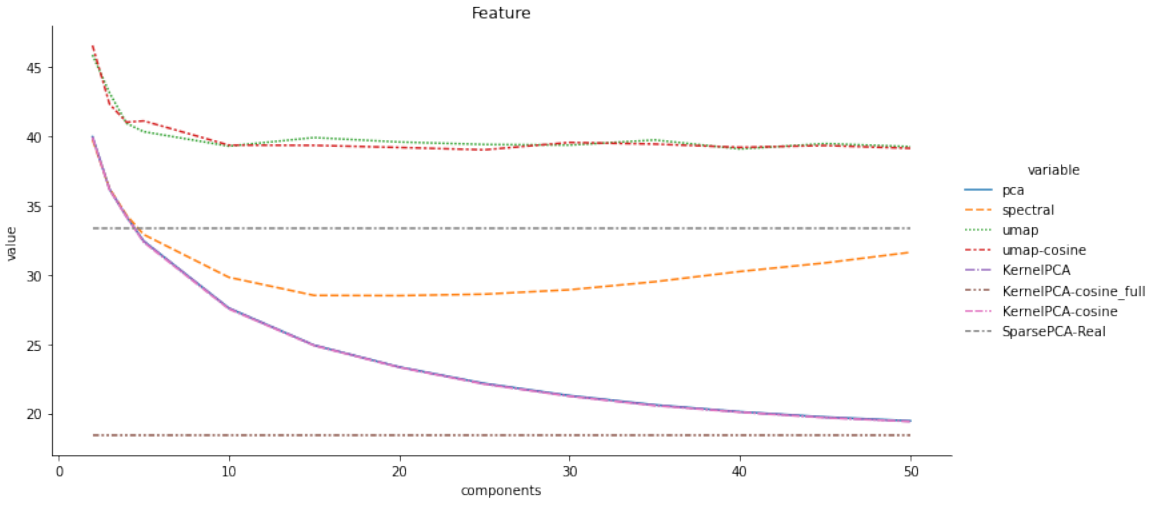

Figure 4-3: Dimensionality Reduction Performance for "Feature" Code Corpus

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Finding Trends in Completed Investigations

Identifying trends in in-service investigations relies on the ability to separate investigations into groups, where each group represents some common feature; in this use case, that feature is the "quality topic." We consider two methods to obtain these groupings: an unsupervised clustering method and a supervised classification method. Unsupervised clustering requires less preparation of ISI data to use, but we find it does not produce groups of quality topics with the homogeneity needed for topic trending. Supervised classification, while requiring data to be properly labeled during training, shows strong performance in labeling investigations with the correct quality topic.

5.1 Unsupervised & Supervised Learning

Unsupervised learning models differ from supervised models in that no "answer" is provided to the model during training. Whereas supervised methods are trained to predict a defined outcome for each input, unsupervised methods instead generate outcomes based solely on the values and variations of their inputs. In this example, this difference manifests as the currently non-existent "quality topic" label for each investigation: a supervised method requires a quality topic to be pre-defined for each investigation, while an unsupervised method does not.

Avoiding the inspection and labeling of historic investigations with quality topic labels is a strong efficiency benefit for an unsupervised clustering approach. Success with this method allows data to be used as-is, generating value without manual intervention. This does require, however, that produced clusters are highly (if not completely) homogeneous. This refers to the model's accuracy in separating investigations by their quality topics - each group of investigations is expected to contain examples of only one type of quality topic as well as every example of that quality topic. Failing to produced homogeneous groups erodes the model's usefulness in extracting informational value, as a quality topic's trend cannot be constructed from the members of a single group. Determining homogeneity exposes a challenge with unsupervised clustering - cluster labeling. Unsupervised methods typically do not generate a descriptive label for the clusters they produce; without this label or any labeling of investigations, ensuring that produced clusters are homogeneous with respect to quality topic falls back on human effort.

Performance of supervised classification, in contrast, is easier to evaluate. As all examples in a data set have labels, mathematical scores of "accuracy," "precision," and "recall" can be easily calculated for a model's predictions. This need for pre-defined labels influences the choice between supervised learning and unsupervised learning. In this scenario, where quality topic labels are absent, a representative sample of the 6,100 investigations would need to be examined by hand and labeled (alternatively, some existing feature could be used as a proxy label). Supervised classifiers generally have poor tolerance for examples of unknown classes - an emerging quality topic that was not seen during training may be mislabeled by a supervised classifier, whereas an unsupervised clustering model naturally acts on this new variation.

5.2 Unsupervised Clustering

The theory that unsupervised clustering will group investigations by quality topic comes from the idea that a person reading a set of investigations (or just their titles and summarizing labels) should be able to determine which ones are similar and

which ones are dissimilar. Based on these similarities, that person should then be able to place investigations into different groups; as the information used for making these comparisons describes the subject of each investigation, these groups would naturally consist of topically similar examples. In automating this process, embeddings that maintain semantic similarity represent each investigation, and the unsupervised clustering algorithm replaces the person creating groups by hand.

Method	**Hierarchical**
Affinity Propagation	N
DBSCAN	N
HDBSCAN	Y
Agglomerative Clustering	Y

Table 5.1: Candidate Unsupervised Clustering Methods

Reduced SBERT embeddings of each investigation's "title," "found" and "produced" codes are used as features in this method. The description given in each title is found to be quite informative with regards to ascertaining an investigation's quality topic; "found" and "produced" codes are included for their value as summarizing labels. We consider four unsupervised clustering methods (Table 5.1) [5][25]. These methods are chosen for their ability to self-determine the proper number of clusters; this allows us to use these methods without needing to know or estimate how many quality topics are represented in the data. Some methods have the additional capability of creating hierarchical clusters; this is potentially advantageous, as quality topics can be hierarchical (for example, "wire knicked" and "wire mislanded" are both more-specific children of the parent "wiring" quality topic).

5.2.1 Process

The Unsupervised Clustering pipeline is comprised of four steps:

1. **Data Selection:**

   (a) **Samples:** A subset of investigations are selected on the basis of gross similarity, such as airplane model, quality issue location, and year reported

(including all investigations is also a valid choice).

(b) **Features:** Features in the subset are downselected to include only "investigation title," "found feature," "found description," "produced feature," and "produced description." [1]

2. **Sentence Embedding:** A transformer is applied to each feature value, transforming text into an embedding representation. This expands the size of the data from $n \times 5$ to $n \times 3840$.

3. **Dimensionality Reduction and Scaling:**

   (a) **Dimensionality Reduction:** Kernel PCA (cosine kernel, $n = 25$) dimensionality reduction is applied iteratively to each original feature's embedded representation, so that only intra-feature variations are considered when creating a reduction. This reduces the size of the data from $n \times 3840$ to $n \times 125$.

   (b) **Scaling:** Standard scaling ($\mu = 0, \sigma = 1$) the resulting reduced embedded representations is experimented with in some trials. Often, no post-scaling is performed.

4. **Clustering:** A clustering algorithm is trained on the final $n \times 125$ reduced representation; predicted cluster membership labels are joined to the pre-embedding text data and returned as the final result.

5.2.2 Results

The unsupervised clustering method ultimately fails to produce persuasive, topically homogeneous clusters of in-service investigations. Some combinations of dimensionality reduction methods and clustering algorithms do provide better clusters, but no combination of techniques yields exclusive groupings of single quality topics: clusters

[1] Identifying data, such as the ISI identifier, is kept to match results back to investigations but is not used in clustering.

either contain members sharing more than one quality topic, produce more than one cluster per quality topic, or exhibit both.

Sample	Component	Quality Topic
1	A	1
2	B	2
3	J	1
4	C	1
5	D	1
6	A	1
7	A	1
8	E	8
9	F	1
10	F	3
11	G	3
12	E	4
13	C	5
14	H	3
15	F	1
16	I	6
17	E	3
18	C	7
19	G	1

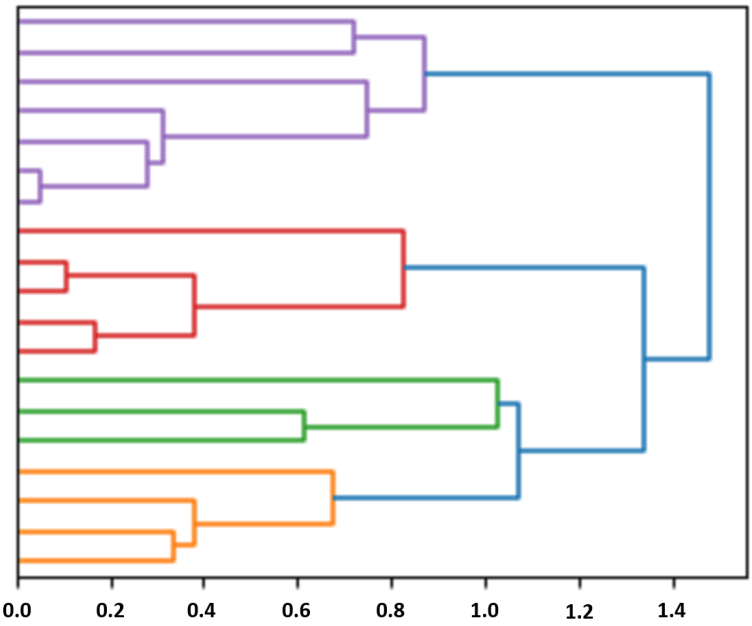

Figure 5-1: Sample Agglomerative Clustering Result - Kernel PCA Reduction (Cosine Kernel, n = 25)

Figure 5-1 shows an example result of performing the unsupervised clustering pipeline on a subset containing 19 investigations. This subset shares a general theme (such as "paint") but represents several distinct types of quality topics (such as "peeling," "bubbling," and "erosion") and different subject components (such as "vertical stabilizer," "fuselage," and "fan cowl"). Agglomerative clustering is used in this example, which attempts to place samples into a hierarchical structure by recursively joining clusters in a way that minimizes the variance of the new cluster. Each sample is listed on the y-axis and "linkage distance" (representing the cluster variance) is listed on the x-axis; each vertical bar represents the creation of a new cluster out of the two clusters (or single samples, if reaching back to the y-axis) that are connected to it. The table to the left of the y-axis lists an enumeration for the component and quality topic each investigation refers to.

This example shows a typical outcome we see in unsupervised clustering. Both clusters of similar components (such as the cluster colored red, which contains most

instances of component "E" and component "F") and clusters of similar quality topics (such as the cluster colored purple, which contains most instances of quality topic "1") are formed, but in neither case are clusters homogeneous. This outcome is an unreliable extractor of informational value as, for any given cluster, human intervention is needed to inspect other groups and confirm that no examples have been mislabeled. Though the accuracy may not seem exceptionally poor from this example, the result given is in reality no more insightful than a keyword search. Similar non-homogeneous clustering performance is seen in trials that are more explicit in their differences in quality topics (for example, a set of "electrical," "fastener," and "paint" investigations), even when the number of samples is small.

A possible explanation for this low level of performance is related to how we represent each investigation as a concatenation of different embeddings. This issue comes from the expanded dimensionality of the feature space, where we represent what was originally five features as 125 columns. The clustering algorithm, however, has no understanding that each set of 25 features (representing a single "native" feature) should be treated as a group - it instead considers them holistically and may attempt to minimize variance in, say, the 20^{th} through 30^{th} columns in a cluster, even though these columns are unrelated in the original space. Reducing the dimensionality of these transformed features to smaller values is a way to minimize this type of problem but, as shown in Section 4.4, lower dimensionality significantly erodes the mutual similarities encoded in the original transformation.

Given these results in small-scale testing, we do not pursue unsupervised clustering for larger scale investigation.

5.3 Supervised Classification

Supervised classification helps bolster the structure of in-service investigation records by adding a new label: quality topic. This label, when added to existing investigation metadata, enables broad quality-centric trending of investigations; without it, human effort is needed to search for examples of each individual topic. Training such

models does come with some effort: training data must be labeled. In this example, ISI records do not contain an explicit quality topic label; rather than perform this labeling step by hand, we choose to use a concatenation of each investigation's "found" feature codes as a proxy for quality topic.

When representing ISI records as features, we examine several options. First is the reduced SBERT representation we use in unsupervised clustering, which preserves semantic similarities in its representations. Second are word frequency-based features, including term-frequency inverse document-frequency (tf-idf). We select these alternative methods under the hypothesis that certain collection of words may be consistently used to describe a quality topic (such as "reversed," "reverse," "swap," "swapped," "wire" to describe a swapped wire).

We examine four types of feature transformations:

1. **Sentence Embedding:** Features are transformed into an embedding space by use of an embedding transformer. The resulting embeddings are reduced using Kernel PCA (cosine kernel, $n = 25$) and, in some instances, standard scaled to $\mu = 0, \sigma = 1$.

2. **Word Count:** Features are passed through a word count vectorizer, which identifies the individual words that comprise a set of examples and returns a matrix listing the number of times each word appears in each example.

3. **Word Existence:** Similar to word count, except array values are set to 1 if a word exists in an example and 0 otherwise. This array now notes only if a word exists in a sample, but does not convey how much the word was used; this representation makes all uses of a word equal, preventing a classifier from being influenced by the number of times a word is used in a sample.

4. **Term-Frequency Inverse Document-Frequency (tf-idf):** Features are passed through a tf-idf vectorizer, which returns a matrix listing a "meaningful importance" value for each word that appears in a sample. These values are calculated to convey the assumption that words that appear less frequently

contain more information - as such, words that appear in few documents are valued higher.

We evaluate classification performance on three types of supervised classification models and several hyperparameter permutations (Table 5.2) [2][4][3]. "Class Balancing" refers to a parameter that re-weighs samples inversely proportional to its classes' occurrence in the training data, attempting to maintain the classifier's predictive power in identifying examples from under-represented classes.

Method	**Kernel**	**Penalty**	**Class Balancing**
Logistic Regression	-	ℓ_1, ℓ_2	Y
Support Vector Machine (SVM)	linear, RBF	-	Y
Random Forrest	-	-	Y

Table 5.2: Candidate Supervised Classification Methods

Performance of each model is evaluated by metrics of precision and recall (Figure 5-2). Precision measures the percentage of times the method is correct in assigning the positive class to an example; recall measures the percentage of examples of the positive class the method is able to correctly identify. This pair of metrics emphasizes the classifier's ability to correctly find and label only true examples of the positive class.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Figure 5-2: Equations of Precision and Recall. "TP" represents true positives, "FP" represents false positives, and "FN" represents false negatives.

5.3.1 Process

Each combinations of feature transformation and classification methods is tested against a hand-constructed data set representing a single investigation "location" (such

as "electrical" or "interiors"). For each set, a positive class is selected (such as "crossed wire") and examples of that class are identified by manual inspection of the data; the set's remaining examples comprise the negative class. Positive class examples typically make up 30% of each data set.

For each test data set, we perform evaluation in four steps:

1. **Data Transformation:** Extract the "title" feature from the data set and apply a given transformation; combine the result with the list of class labels.

2. **Build Classifiers:** Construct a list of classifier instances which cover all hyperparameter combinations (kernel function, weight balancing, and standard scaling) of interest.

3. **5-Fold Cross Validation:** For each classifier instance, perform 5-fold cross validation on test data. Prediction performance on the "test" data segment (precision and recall metrics) are recorded for each trial.

4. **Performance Reporting:** Consolidate and plot individual classifier performance.

Note that we only consider investigation titles as features in this evaluation and not "found" or "produced" codes. This is because we now use these codes as a proxy for quality topic; including their values in the input would produce inaccurate results.

5.3.2 Results

Strong classification performance is observed in trials of both embedding-based and word-based feature representations. Embedding features generally show a better composite performance of precision and recall: Figure 5-3 and Figure 5-4 (showing classifiers trained on 50% of the ISI data set to predict "crossed wire") show support vector machine (SVM) classifiers performing above 80% on correctly labeling the positive class and above 75% performance on correctly identifying existing instances of the positive class. Logistic regression performed less-well, with precision and recall typically at or below 40%. Random forest classifiers do outperform SVM in measures

of precision (labeling the positive class correctly 90% of the time), but suffer in recall (identifying less than 40% of examples of the positive class).

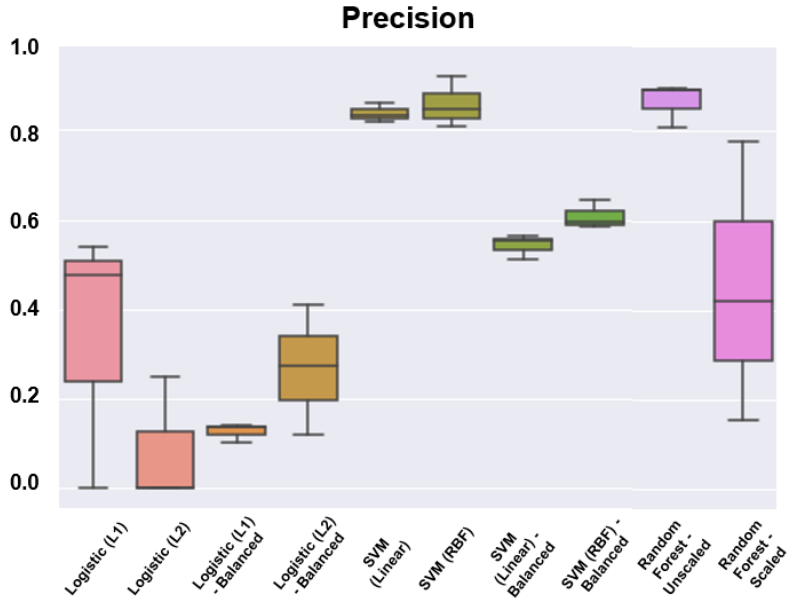

Figure 5-3: Code Prediction Classifier Result (Precision) Predicting "Crossed Wire" from Investigation Title Sentence Embedding Transformation (50% Sample)

	Logistic Regression (ℓ_1)	**SVM (Linear)**	**SVM (RBF)**
Precision	15%	81%	90%
Recall	10%	79%	75%

Table 5.3: Code Prediction Classifier Result. Predicting "Crossed Wire" from Investigation Title Sentence Embedding (100% Sample)

Expanding the training data set to include all ISI examples raises both precision and recall absolute performance and tightens variance, with SVM classifiers achieving best results (Figure 5-5 and Table 5.3).

Classifiers assessed with word-based feature transformations showed lower performance than their sentence embedding counterparts. Figure 5-6 and Figure 5-7 show results of classifiers trained on word count feature representations - as with the embedding example, classifiers are trained on 50% of the ISI data set to predict "crossed wires." While precision performance of logistic regression classifiers is high (at or above 80%), recall is very low (at or below 30%). SVM classifiers, which perform well with embedding representations, fail to consistently show recall above the 60%

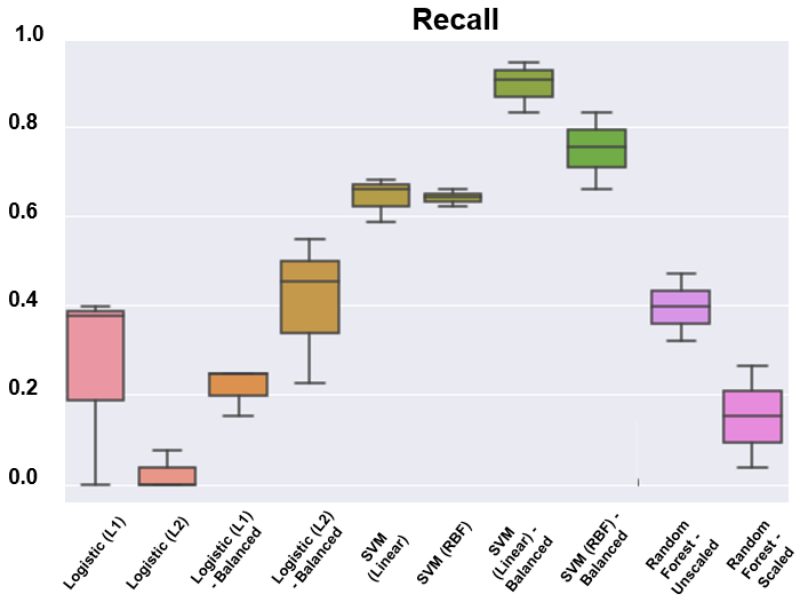

Figure 5-4: Code Prediction Classifier Result (Recall) Predicting "Crossed Wire" from Investigation Title Sentence Embedding Transformation (50% Sample)

threshold. Random forest models again show great precision (100%, in this instance) but very low recall (10%).

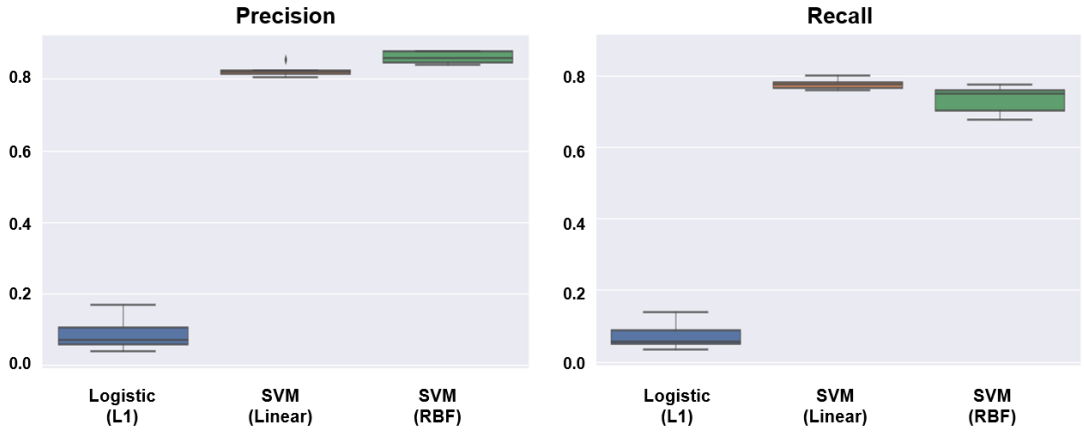

Figure 5-5: Code Prediction Classifier Result Predicting "Crossed Wire" from Investigation Title Sentence Embedding (100% Sample)

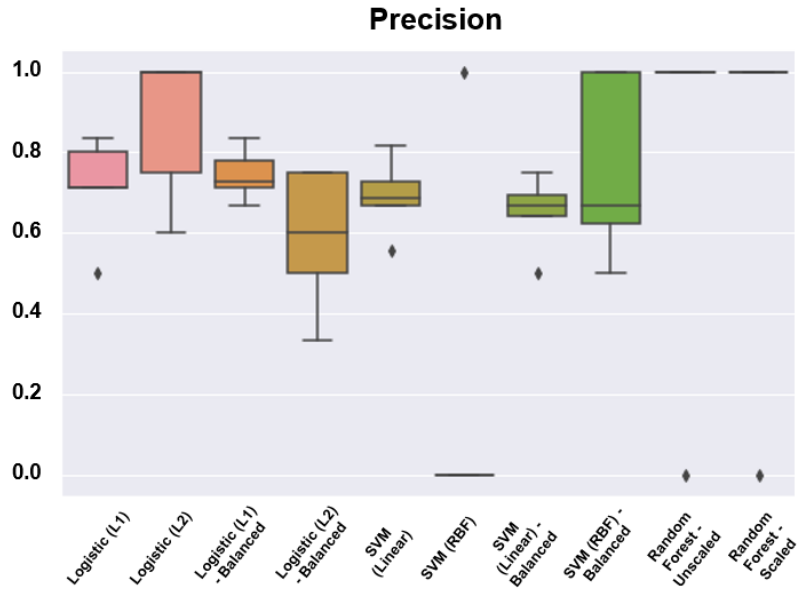

Figure 5-6: Code Prediction Classifier Result (Precision) Predicting "Crossed Wire" from Investigation Title Word Count Transformation (50% Sample)

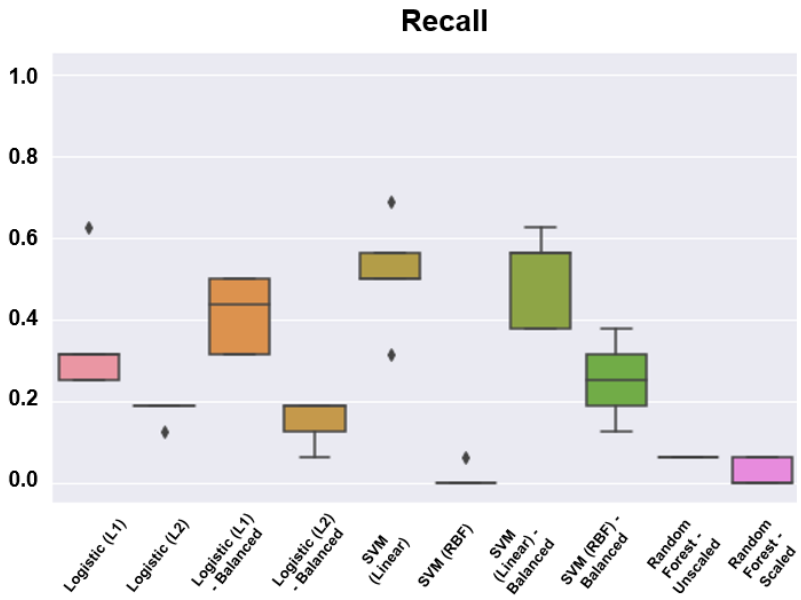

Figure 5-7: Code Prediction Classifier Result (Recall) Predicting "Crossed Wire" from Investigation Title Word Count Transformation (50% Sample)

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Improving Investigation Performance

A significant portion of the ISI workflow involves the act of conducting new investigations. This includes exploring quality records for instances that either relate to (i.e., similar occurrence) or may have contributed to the current investigation; after an investigator finds potentially relevant documents, they must judge which are related to the current investigation. As both of these parts are manual, outcomes vary based on the investigator's skill and experience. We look at a way to add standardization to this process through use of an investigation "helper" tool: after automatically loading the relevant data, this tool uses a variety of similarity metrics to assist investigators in comparing both structured and free-text data, improving the efficiency and consistency of finding records of interest.

6.1 NLP-Based Investigation "Helper"

Each in-service investigation includes a review of, among other sources, past investigations for instances of similar quality occurrences - as each investigation concludes with corrective actions, seeing a quality issue repeat may point to previous corrections being inadequate. While identifying these trends in quality topics without significant human intervention is currently unavailable (hence its focus in this thesis), investigators must regularly expend this effort on each investigation to both collect and analyze potentially related documents. One of the investigator's main tools in this

process is the keyword search: using their knowledge and experience of how quality issues are described, they will query with different words and phrases in an attempt to find similar investigations. Another tool is the part number search - investigators will use the datum part number as a query term to find other investigations involving that part, much in the same way as a keyword search.

Notice that these types of searches attempt to identify documents through similarities in actual word choice or semantic meaning, similar to the type of representations sought in Chapter 5. We attempt to apply the same text-transforming methods here, presenting the investigator with the similarity information directly. Such information improves the efficiency and completeness of each investigation: searches no longer rely solely on each investigator's prowess to collect or relate documents, leading to more consistent results. We choose comparison of only in-service investigations for this example, as it is well explored in this thesis. Inclusion of additional data sources examined in an investigation, including production and delivery records, is an additional opportunity.

6.2 Part Number N-Gram

While we have the ability to ascertain similarities between written prose using our embedding methods, we lack a similar capability to effectively compare free-text representation of part numbers. Within the ISI data set, identical part numbers rarely appear; what is more common, however, is to see repeated root part numbers (e.g., "W12-11A" and "W12-12A" share "W12"). As some part numbers are hierarchical (such as wire labeling, where part numbers include both the wire bundle number and the individual conductor label) as seen in Figure 6-1, it is advantageous to have a comparison technique that also considers similarity in these roots. Extracting these part numbers is complicated by their representation as free-text - we often find shorthand representations when multiple part numbers are listed (e.g., "W12-11/12A" to abbreviate "W12-11A & W12-12A"). The ability to expand these representations into full part numbers and create a more structured representation improves the

completeness of our similarity comparison efforts.

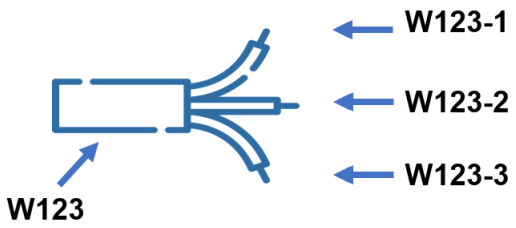

Figure 6-1: Hierarchical Wire Bundle and Conductor Part Numbering

Step	Transformation	Example A	Example B
0	Original Representation	"Paint, W12, (W13-001/-002)"	"FOD , (W13-001-1A)"
1	Remove Brackets and Split on Whitespace	["Paint", "W12", "W13-001/-002"]	["FOD", "W13-001-1A"]
2	Remove any all-letter items	["W12", "W13-001/-002"]	["W13-001-1A"]
3	Expand "-a/-b" shorthand	["W12", "W13-001", "W13-002"]	["W13-001-1A"]
4	Recursively split on "-" to create N-Grams	["W12", "W13", "W13-001", "W13-002"]	["W13", "W13-001", "W13-001-1A"]

Table 6.1: Part N-Gram Algorithm and Example

	W12	W13	W13-001	W13-002	W13-001-1A
Example A	·	·	·	·	
Example B		·	·		·

Table 6.2: Part N-Gram Sparse Matrix Representation

We devise an algorithm called "part n-gram"[1] to perform this expansion and re-structuring of free-text part number metadata (Table 6.1). This algorithm is rule based, applying a heuristic developed through manual inspection of ISI part number metadata to expand commonly used types of shorthand before recursively splitting hierarchical part numbers into a set of potential roots. We apply part n-gram as a

[1]"n-gram" is inspired by the computational linguistics n-gram process, where a text sample is broken into a set of each n contiguous words. However, this method does not pre-define n.

Dot Product	Cosine Similarity
2	0.58

Table 6.3: Part N-Gram Similarity Comparison

transformation, changing a list of part numbers described in free-text into a sparse matrix representation (Table 6.2).

Comparing part numbers between two examples in this representation is simply computed as the dot product of their respective vectors: a value of 0 indicates no part numbers or part number roots in common, while a value above 0 indicates some shared part number root that can be investigated further (with larger values indicating a higher number of shared components). Normalizing this dot product by the magnitude of the resulting vectors yields the cosine similarity of the examples, bounding the similarity value between 0 (no parts in common) and 1 (identical part numbers), as shown in Table 6.3 for the matrix representation shown in Table 6.2.

6.3 Use Case Example

We demonstrate this helper tool's use as an investigative aid through the following example. While the data shown is fictitious, scores listed in each wire frame are computed using the similarity methods described (SBERT text embedding and part n-gram transformation) and portray actual performance.

Figure 6-2 shows the starting point for a newly opened investigation: the investigation is given a title ("Incorrect Bin Stencil Found on Airplane"), the involved part number is identified ("400100-21#A"), and the condition in which the quality issue was found is coded ("stencil," "not per drawing").

One first step for an investigator may be to search for any other investigations that involve the same or similar part numbers; such investigations may point to a systemic issue, or to a corrective measure that has failed to completely fix the problem. Investigations that include multiple part numbers make this search even more time consuming: each part number (and its root, in the case of hierarchical part numbers) must be searched for individually. Our tool simplifies this workflow by performing such

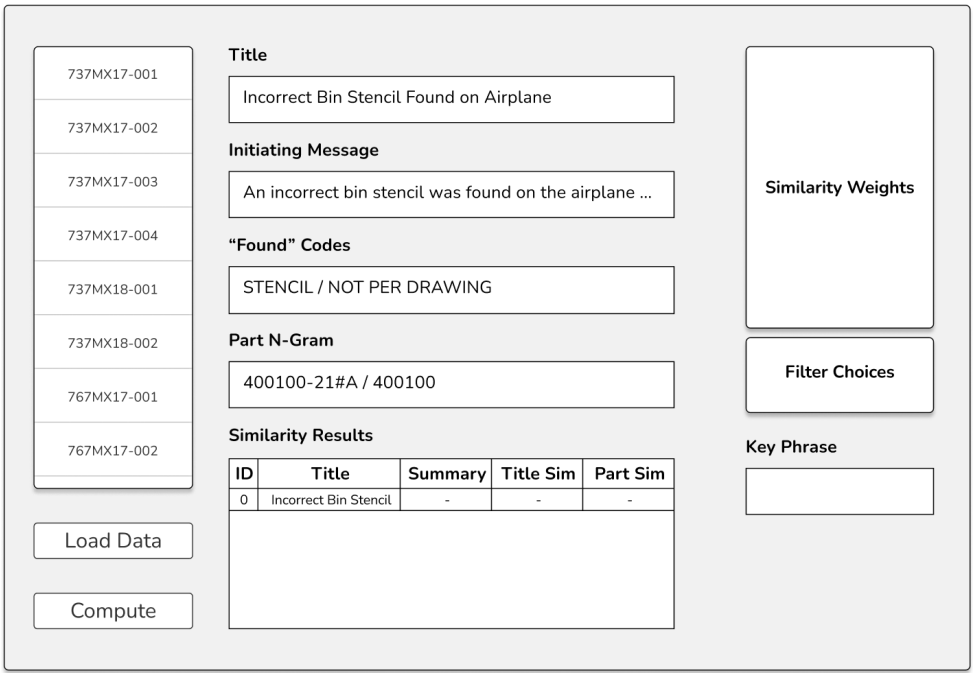

Figure 6-2: Investigation Helper - Datum Investigation Selected

a search automatically - each investigation's part number metadata is transformed by part n-gram and compared, allowing for a single similarity score to be returned. Figure 6-3 shows this result: an investigation that shares the same root part number ("400100") is found through a non-zero part similarity score.

A next step may be to search for investigations that have similar titles. Our review of ISI records shows that these titles are typically high quality with respect to summarizing the investigation's quality topic. Searching for similar titles typically involves a keyword search: investigators decide which words are germane to the quality topic, if there are other ways to describe the issue that should also be queried, and if there are abbreviations that should be included in a search as well. Our tool simplifies this by process by computing similarity scores from each title's embedding representation. Figure 6-4 shows such a comparison: an investigation titled "Incorrect / Missing Stencil" has a strong title similarity score to the datum investigation. Note the absence of the word "bin" in this title - a strict keyword search for "bin stencil" (as the problem is described in the datum investigation) would not find this investigation.

The ability of our chosen embedding transformations to find semantic similarity

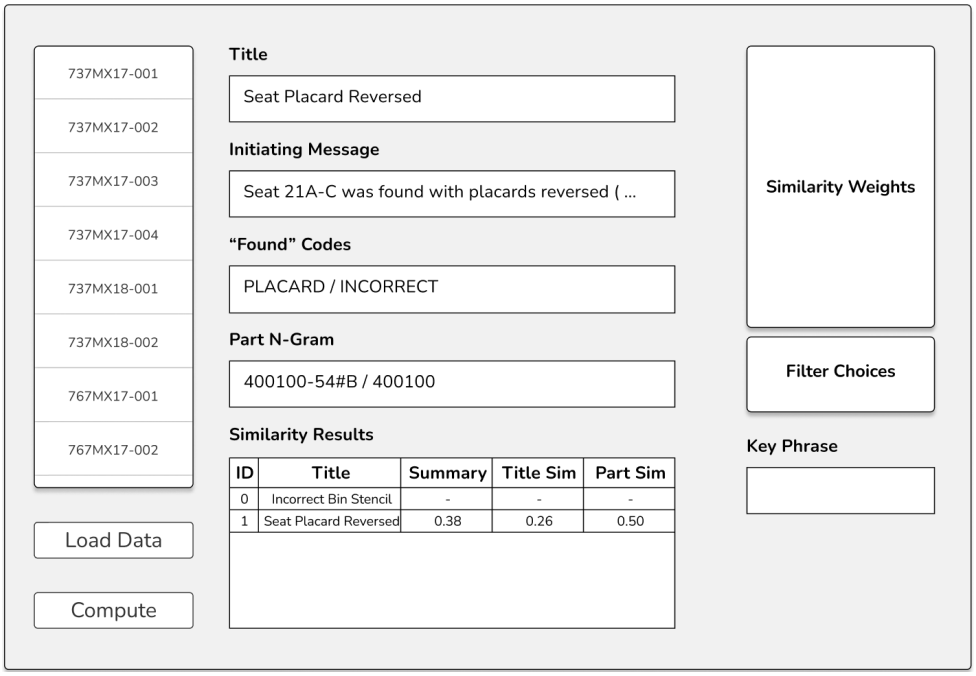

Figure 6-3: Investigation Helper - Similar Part Number to Datum Investigation

Title	Similarity
Incorrect Stow Bin Stencil	-
Overhead Stowage Bin Stencil Missing	0.81
Stow Bin Hinge Bracket Broken	0.62
Loose Overhead Bins	0.39

Table 6.4: Investigation Helper - Sentence Similarity Finding Alternative Phrases

reduces the need for investigators to consider other ways of describing a given quality topic. Table 6.4 lists three example investigation titles and their similarity scores to "incorrect stow bin stencil." In the first example, "stowage bin" is alternative wording for "stow bin" - similarities between the phrases "incorrect stencil" and "stencil missing" improve its score. Again, a strict keyword search for "stow bin" would likely not have found this result.

6.4 Results

This investigation helper, in its current form, exists more as a conceptual prototype than as a production-ready tool. It is built as an interactive demonstration of how

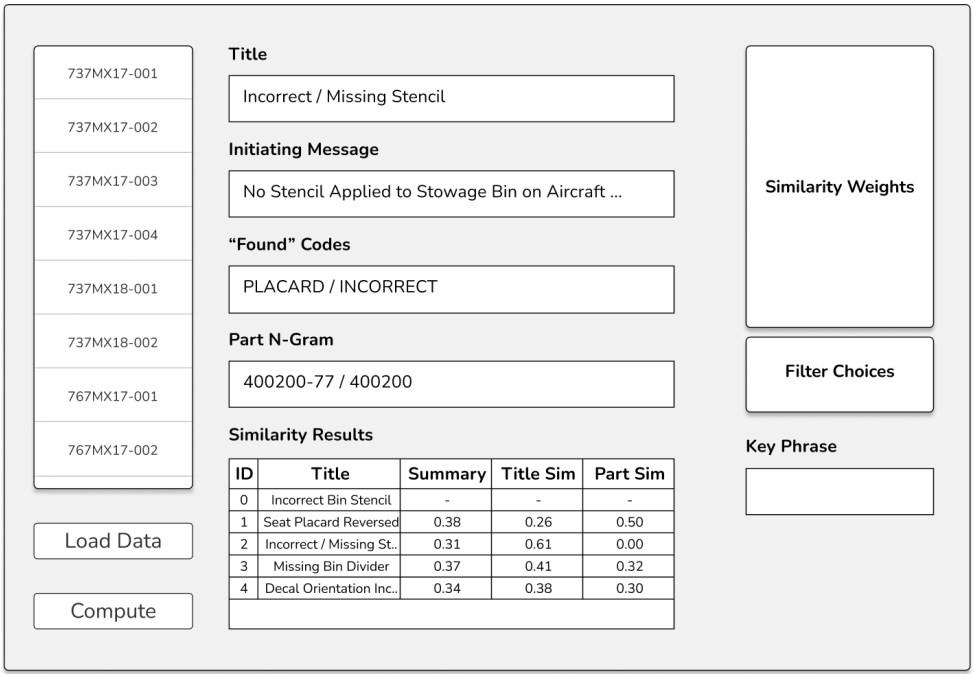

Figure 6-4: Investigation Helper - Similar Titles to Datum Investigation

data extracting methods (both the part n-gram algorithm and NLP-based similarity metrics) can benefit online processes, such as comparing an arbitrary set of in-service investigations. It also demonstrates a fairly trivial point of data access: it is easier to see things in one view. In the case of the in-service investigation data set, different features reside in different data stores; manual effort is required to collect and merge these results together into the view presented.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 7

Summary, Future Work, & Recommendations

In Section 7.1, we summarize the results of our attempts to extract informational data from unstructured, free-text in-service investigation reports using a mixture of NLP-based embedding methods and machine learning techniques. Most promising success comes from supervised classification, as trained classifiers have the ability to automatically assign a quality topic label to both new and old investigations. This new piece of metadata enables grouping investigations into broader classes of quality issues, as well as aligns the nomenclature used to describe given quality issues with that used in other parts of the commercial airplane value stream. We further extend our use of similarity-finding methods in an investigation "helper" tool, discussed in Section 7.2; this tool, crafted as a demonstration, automates both the data collection and mutual comparison processes investigators must complete when reviewing in-service investigation records.

We review the choice of the in-service investigation process as an appropriate use case in Section 7.3, and find positive support from the types of quality issues investigated and the organizational need for more accessible reporting regarding these issues. In Section 7.4, we recommend several areas for improvement that can be enabled by our findings: increasing consistency of quality issue descriptions within the value stream, emphasizing completing these descriptions as information becomes

available, and considering a shift to a product-centric approach over a project-centric approach when implementing these methods.

We conclude in Section 7.5 with a perspective on the sentiment towards proactive quality and machine learning enablers in broader domains. We are encouraged by both active industrial and regulatory interest in these areas, as well as growth in commercial products offering capabilities similar to what we describe in this work.

7.1 Extracting Information from Completed Investigations

We explore two distinct methods for extracting information from the largely free-text in-service investigation documents. In both cases, we attempt to group these documents by the "quality topics" they represent. The ability to separate investigations into these groups supports analysis and proactive action, as it enables understanding of how frequent a given type of quality issue is reported and what similarities may exist between these occurrences (Figure 7-1).

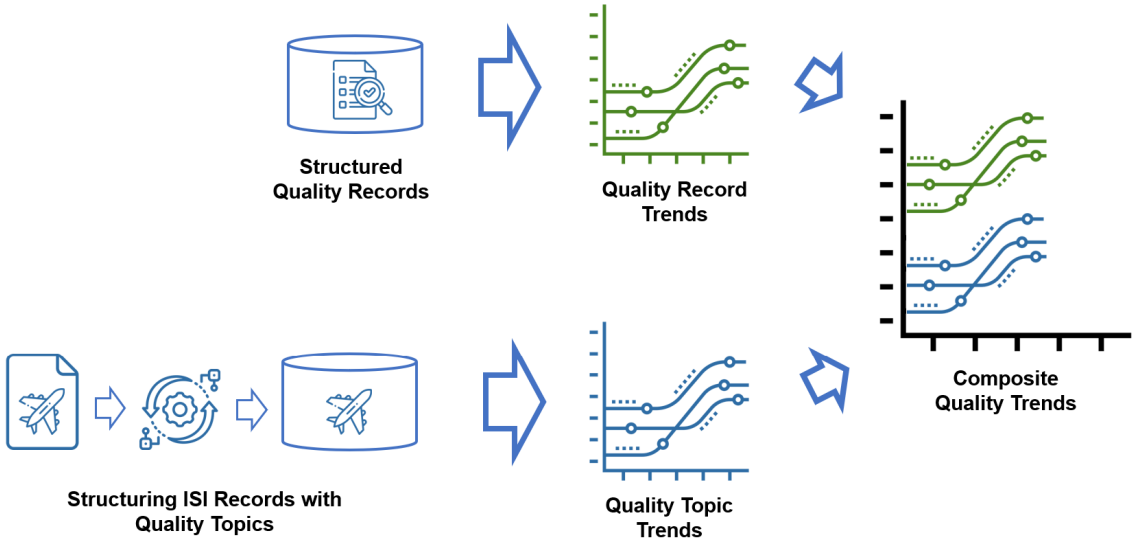

Figure 7-1: Combining Quality Topic Trends with Existing Quality Trends

The first method we explore is an unsupervised clustering algorithm, which we train to separate investigations into groups that each represent a single quality topic.

The second is a supervised classification algorithm, which we use to directly assign quality topic labels to investigations. Supporting these methods are several natural language processing techniques that aid us in representing free text data as something more numerical, which is required by our learning algorithms.

7.1.1 Unsupervised Clustering Models

Unsupervised clustering models are attractive for this use case due to their flexibility and low "barrier to entry" for data organization. Flexibility comes from the ability to self-determine the best number of clusters that fit the input data. While this is not a capability of all unsupervised algorithms, it is of the algorithms we choose to explore. This is beneficial from the standpoint of where we are with our unstructured data - we are usually not sure what exactly is in it. Should we have chosen an algorithm that did require a pre-defined number of clusters, we would need to manually inspect our data set to determine the number of quality topics it contains; this effort runs counter to our desire for automatic clustering. Similarly, unsupervised methods do not require any label or sense of "truth" in training or evaluating data, and instead operate on the variations observed within the data set. This again helps us avoid manual inspection or re-processing of data to contain a required feature - we can use both historic and contemporary data alike, almost immediately.

Performance of our unsupervised clustering models, even in small-scale testing, is not to the level required to enable quality topic analysis. While generated clusters do generally contain examples of a single quality topic, they are rarely homogeneous or complete. Features we choose for this method (investigation title, "found" and "produced" summary codes) have high informational value with regards to each investigation's quality topic, but were not sufficient to obtain complete separation of investigations. The question of how to label produced groups as a given quality topic is acknowledged but not yet solved; human effort may be required to post-process each group and assign a label. A more fine-tuned embedding model, trained on data akin to what is found in in-service investigations, may improve performance for this method to a degree where useful information is produced.

7.1.2 Supervised Classification Models

Supervised classification methods require more up-front thought when training a new model: data needs to be properly labeled, and training sets need to be constructed in a way that represent the variety and composition of data that the model is expected to see in practice. For use cases like ours, where some preparation to find and label training data is needed, this can require a large amount of effort. One benefit of this type of supervised method (compared to the unsupervised method) is that result structure is more consistent - model outputs will always be one of a known set of class labels.

Classifiers we train to predict an investigation's quality topic (using a concatenation of "found" codes as a proxy for this label) perform very well. Support vector machine models, for example, reach above 80% precision and above 75% recall in prediction performance. Input features comprised of semantic-free word frequency and semantic-maintaining embedding transformations both perform well, with embedding representations showing best results. As these embeddings are generated from a pre-trained model that has not been specifically exposed to Boeing or aviation industry corpus, even better performance may be obtained with a more purpose-built transformer model. While our exploration focuses on using transformations of only an investigation's title as the input feature, different ways of representing an investigation may also yield better results. Boeing is researching methods that can extract quality conditions from text documents [26] - such information has high value for classifying quality topics and may make for strong features.

Despite the additional effort required to train classification models, the combination of performance and structure we observe in results makes this method a strong candidate for extracting information from unstructured data, both historical and newly-created: these classifiers can both assign quality topics to completed investigations without the need for detailed human review, and assist investigators in choosing the correct quality topic for new investigations. The proactive quality processes we anticipate integrating this data with rely on structured representations - having a

known set of quality topics adheres to this requirement much better than the result of the unsupervised method (which produces a varying number of un-labeled clusters).

7.2 Improving New Investigations

We devise an investigation "helper" tool to serve two purposes: act as a discussion piece on an alternative investigation workflow, and provide a more direct demonstration of similarity-finding with NLP methods.

This tool implements a new way for investigators to gather information during the data discovery phase of a new investigation. While the steps to conduct an investigation are defined, the act of effectively gathering data relies on the investigator's mastery of internal tools. Different types of quality data (test records from production and flight squawks from delivery, for example) are accessed through different front-end views (Figure 7-2): each investigator has to know the best way to query that view (in some instances, pre-defined filters that are informally passed down; in others, effective use of wildcard searches), which impacts the completeness of data returned. Reports summarizing these found quality documents are often represented as separate spreadsheets - investigators either compare these individually or attempt to create their own summarizations, taking time and adding opportunities for error. These front-end views, however, get their data from a centralized "data lake" which can be directly queried. With this tool, we propose exploiting this capability and automating the document collection aspect of data discovery. Doing so will add consistency in both completeness, as all searches will use a common method, and presentation, as the tool performs appropriate merge and join operations on its intermediate results.

This new standardized presentation enables our tool's similarity-finding capabilities. We automate two common methods of comparison: semantic similarity (finding uses of the same or similar words), and part number similarity (finding instances of the same or similar part numbers). Semantic similarity comparisons are computed using the SBERT embedding model, while part number similarity is computed with a bespoke "part n-gram" algorithm which transforms written short-hand description of part

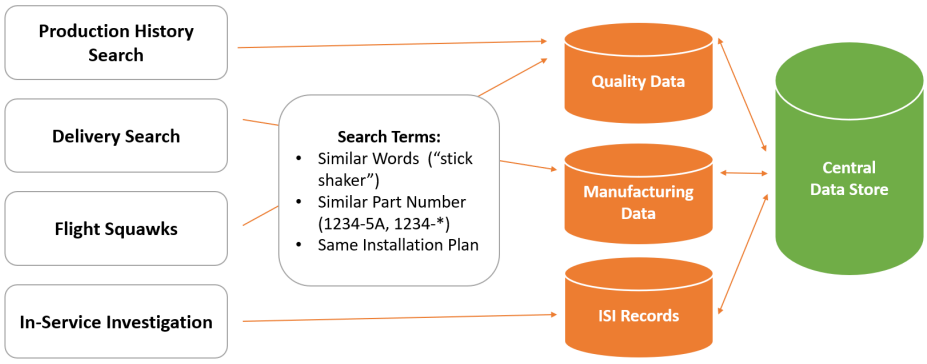

Figure 7-2: In-Service Investigation - Current Search Process

numbers into an expanded representation. In production use, we anticipate these methods will improve the speed and completeness of finding related quality documents: investigators will be guided to relevant records by these similarity scores without the need for manual searching. Our implementation displays similarity scores from our comparison methods, rather than a ranking. We anticipate this has educational value - as NLP methods are new to many in traditional quality roles, the ability to see the performance on data they are familiar with helps build understanding of their capabilities.

As a persuasive discussion piece, we find mixed results. Through our demonstrations using this tool, we find parties at various levels in the organization who acknowledge the variability in how efficiently investigations are performed, and agree that such a tool could add consistency and improve results (particularly as, with many BCA functions, investigator teams are split by airplane program). This sentiment is not, however, shared universally; we find other parties who do not believe that such a tool will add value, and instead believe that variability can be best solved through investigator training. Additional effort to refine the premise of this tool and consider the viewpoints of those who question its value is likely needed before formal development is endorsed.

7.3 Reviewing the In-Service Investigation Use Case

The premise of this study looks at how unstructured data can be manipulated or mined in ways that allow its informational value to be used on par with data

that comes in more structured forms. We choose the in-service investigation process as a practical use case to explore. In addition to providing an interesting example of free-text data, the ISI process also has existing internal motivators for reform in reporting and integration with the larger quality management function.

We find summary reporting to be a strong area of improvement for the in-service investigation process. Two common questions we find in a large number of the customer messages that initiate an investigation are variations of "has this happened before?" and "is still still happening?"; as there is no consistent summary of what quality escapes have been previously reported, these questions must be answered through repetitive manual reviews of past investigations. These questions reappear in other parts of the value stream. Customers taking delivery of a new aircraft will often ask about quality concerns that they have experienced and reported through the ISI process. As "delivery" is separate from "in-service" in the BCA matrix organization, these questions are even less immediately answerable. We observe leadership acknowledgment of this informational gap, but experience proposed resolutions that rely solely on mining existing metadata and that do not consider changes to the ISI process. We believe the quality topic labeling method we propose will resolve this gap, as investigations will now be grouped into reasonably specific quality categories in a format that is accessible to the broader organization.

We believe categorizing in-service investigations by quality topic enables insightful analysis. By forming these groups of similar quality escapes, we can more easily explore common conditions (noticing, for example, that a particular flow day contributes to the majority of loose fasteners) and identify more systemic issues (finding, for example, that all airplane programs have high numbers of wiring quality issues). Consistent use of this approach also helps us break the organizational airplane program barrier. As investigation data is stored in a consistent format for all investigations, we can apply a single method to examples from all airplane programs and generate a holistic view of the types of quality issues reported through the ISI mechanism.

Our review of past in-service investigations and the types of issues that are reported supports our belief that trends from these investigations are valuable information for

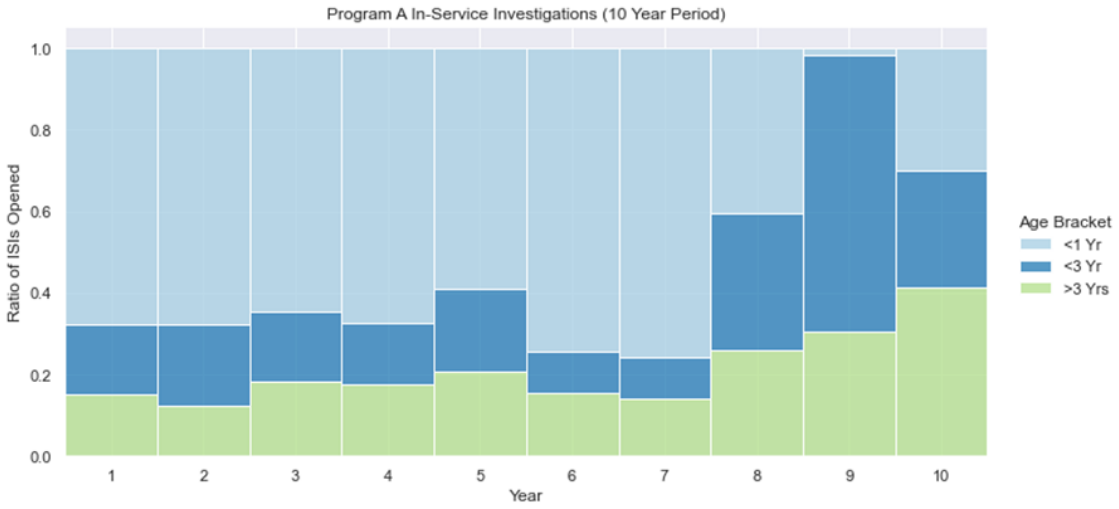

Figure 7-3: In-Service Investigations - Distribution by Airplane Age

current proactive quality practices. Topically, we find that a majority of investigations pertain to quality issues that may be a result of production or delivery practices, rather than airplane "wear and tear." This is supported by the distribution of aircraft ages - a majority of investigations involve aircraft that are less than three years old (Figure 7-3), with over half involving aircraft less than one year old. This suggests that quality trends from in-service investigations are not areas of quality management separate from production or delivery, but are actually lagging indicators of quality performance in these earlier steps of the value stream. Inclusion of ISI information, supported by our quality topic categorization, directly benefits existing proactive quality practices in these areas.

7.4 Boeing-Specific Recommendations

Our work on the in-service investigation use case exposes us to the organizational dynamics and culture within Boeing Commercial Airplanes. While technical performance is necessary for our proposed methods to succeed, the business environment in which they are deployed can have equal impact on adoption. We note three areas with opportunities for organizational improvement: aligning how summarizing codes are used in quality documents, reducing the time to complete these codes, and considering

how the "project" view of work may not be the most resilient in an organization with high movement.

7.4.1 Align Coding Usage

Two pieces of metadata we commonly mention and utilize throughout this thesis are the "found" and "produced" codes associated with each in-service investigation. These standardized codes summarize how a given quality issue presents in the product, and how it was introduced into the product, respectively. As these codes are artifacts of the manufacturing execution system's quality documentation process, they are found in quality reports from production and delivery phases as well. We find, however, that utilization of these codes varies strongly between these value stream phases: in-service investigations are less likely to include complete, descriptive codes.

Interviews with in-service investigation team members reveal that these codes are not emphasized to the degree we see in production or delivery. The closure process focuses on verifying that values are filled in, but does not include guidance on what codes should be used for different quality topics (for example, standardizing on "wire mislocated" or "electrical miswired" to describe an incorrectly wired lighting fixture). Review of the application of these codes shows that correct usage provides the descriptive "quality topic" label we seek - we find through review of historic investigations that only about 30 unique "found" codes are needed to label over 50% of these examples with the appropriate quality topic.

We recommend a standardized approach to labeling in-service investigations with "found" and "produced" codes that is consistent with the labeling of similar quality issues in other areas of the value stream. We believe the supervised classification method we demonstrate will be a strong enabler of this plan: models can be trained to label both new and historic investigations with correct codes, focusing on the 30 values we find to represent a large number of investigations. Standardizing on this summary information improves the ease of relating quality documents originating in different parts of the value stream.

7.4.2 Code Information on Arrival

As summarizing metadata becomes increasingly used for analyzing in-service investigations, the timely application of this information becomes more important. We find that such values are not consistently completed in the early phases of new investigations (when the reported condition is being clarified with the customer and a record is being created); instead, an investigation may have to make it to closure (several weeks to months later) before metadata is fully entered (Figure 7-4). In instances like the "found" codes which describe the reported condition, sufficient information to fill in these values is available before investigation closure.

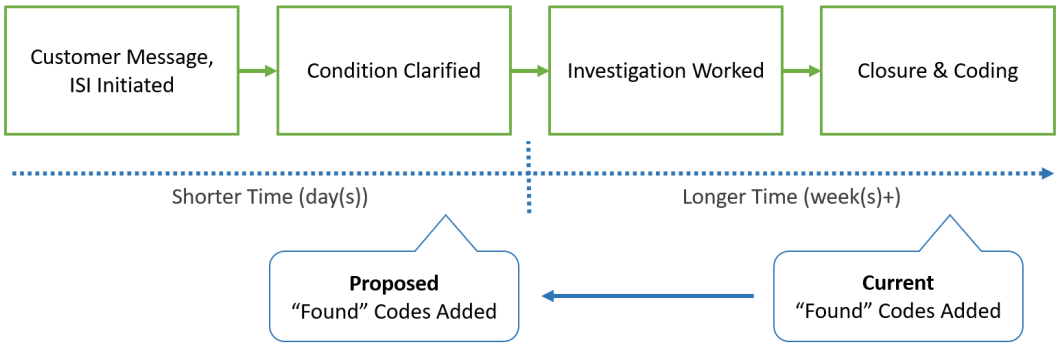

Figure 7-4: In-Service Investigation Workflow: Current and Proposed Coding Steps

We recommend a change to the order of in-service investigation steps to emphasize coding information as it becomes available; delaying the entry of these values delays information becoming available to broader contexts. Completing codes and other metadata values as information becomes available eliminates this lag with no negative impact to the investigation process.

7.4.3 Consider Barriers to Organizational Traction

Organizational considerations can play as much role in a new process' success as technical ability does. Through this study, we experience the nuances of Boeing's organization and its dynamics. We note two areas that show opportunity for improvement with regards to supporting novel changes: consistency of support, and a "project" vs. "product" mindset.

Roles within BCA, particularly in the professional skill sets and management ranks, are characterized by a high amount of flux. Employees move fairly fluidly within the organization, moving to new positions sometimes as frequently as twice within a year. "Temping" is commonly experienced in supervisory positions, where a team leader is temporarily assigned to cover the role until a permanent placement can be found; individual contributor roles are more static in comparison. The result of this frequent movement is a constant cycle of re-engagement and re-education as new people move into stakeholder roles. This presents a barrier to progress, as decisions that depend on these stakeholders are often stalled or recycled as the incumbent becomes familiar with their new role. Avenues to move the engagements needed to develop new workflows such as ours to more static stakeholders would benefit both overall efficiency and success.

Some groups within BCA are experimenting with moving to agile work practices, where work is broken into small but easily-validated pieces, including the quality improvement team who support this project. We find value in the frequent progress meetings that Boeing's implementation of agile host: updates and questions are shared in short bursts twice a week, with longer demonstrations of progress being held approximately quarterly. This improves the feedback cycle and lowers the risk of expending effort in a direction that is unsupported by stakeholders. However, we find Boeing's view of agile to support a project-centric view of work vs. a product-centric view. As a project, our effort seeks buy-in from team leaders and those with organizational power on its direction; we are influenced by the needs and opinions of those within their teams, but can not act without a higher endorsement. This contrasts with a product-centric view, where direction is chosen by the product owner and success is judged by a product's adoption and realized performance. We believe the product view is more compelling in instances such as ours, where we target improvements to performance and "quality of life" for a type of knowledge worker position. We also believe this may balance the observed stakeholder flux: by leaving only very high-level decision approval with these stakeholders, a product-centric approach is more tolerant to changes in such roles.

7.5 Broader Interest in Proactivity, Quality, and Machine Learning

The combination of quality management, proactive processes, and natural language processing we present in this work follows broader trends in both academia and industry. The pursuit of quality in manufacturing is by no means new, nor is the notion that proactively working to solve problems when they are small can be less costly than reactively battling the larger issues they grow in to. Where we see added vigor is in the application of modern machine learning methods to address these challenges: insight and relationship-finding can now come from learning algorithms, reducing the amount of human effort required to inspect and analyze complex data. We note interest in proactivity, quality, and machine learning through two lenses: industry sentiment, and commercial product offerings.

The desirability of proactive methodologies within industry and the benefits they bring is seen in several places. One is in conference proceedings, where different flavors of data-driven proactivity are frequently presented. Oil & gas industry forums, for instance, often share innovations where "quality" comes from proactive operations and proactive maintenance: recent examples include utilizing increasingly exotic instrumentation to support proactive operations decisions with more informative measurements (analyzed by hand in this paper, but providing a foundation for future automated means) [23], and employing machine learning methods on equipment telemetry to create better proactive maintenance triggers [36]. Other sentiment is seen in regulatory messaging. The Food and Drug Administration (FDA) is pushing medical device manufacturers to move past regulatory minimums and "actively seek out quality issues" as a means to curtail increasing instances of serious adverse events, finding necessity in the inclusion of proactive quality practices [1]. Aside from the dominating concern of patient health, the financial impact of improvement is huge: McKinsey estimates the cost of poor quality in the medical device industry to be between $26 billion and $36 billion annually [17].

Machine learning-enabled solutions to industry problems are seen in commercial

products with growing frequency. Sparta Systems, a division of industrial conglomerate Honeywell, has included natural language processing in its line of quality management software focused on the life science industry since 2018 [7]. One such capability, "QualityWise.ai," seeks to implement document classification similar to what we present for in-service investigations: summarizing categorical labels are suggested based on NLP analysis of free-text quality reports [39]. MasterControl, a quality management software company targeting the life science and process industries, uses NLP in a matter reminiscent of our investigation "helper," providing "Netflix-style contextual search and recommendation engines" that aim to "improve search speed and accelerate the retrieval of relevant information" for quality documents [19].

Interest in the applications of new technologies and the availability of "off-the-shelf" implementations support the proliferation of machine learning in industry. While companies are rarely short on potential areas of improvement, pursuing a solution from the variety of options in the machine learning mosaic can be daunting; product vendors help ease this burden by packaging application-relevant technologies in formats that meet these customer needs. This growing marketplace, seeded with work similar to our own, enables a much broader audience to reap the benefits from these methods, helping to move the cutting edge of technology from the academic to the practical.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

SortOrder

```
import scipy.stats
import numpy as np

def sort_order(a,b):
    '''
    Solves the re-ordering of B to match A and
    returns the scaled number of steps required to re-order
    '''
    moves = 0

    while not np.array_equal(a,b):
        lsb = np.argwhere(np.not_equal(a,b)).min()
        msb = np.argwhere(np.not_equal(a,b)).max()

        temp = a[lsb]
        a_pos_1 = int(np.argwhere(a == temp))
        b_pos_1 = int(np.argwhere(b == temp))
        u = np.insert(np.delete(b,b_pos_1),a_pos_1,temp)
        u_met = scipy.stats.kendalltau(a,u)[0]
```

```
        temp = a[msb]
        a_pos_2 = int(np.argwhere(a == temp))
        b_pos_2 = int(np.argwhere(b == temp))
        v = np.insert(np.delete(b,b_pos_2),a_pos_2,temp)
        v_met = scipy.stats.kendalltau(a,v)[0]

        if v_met > u_met:
            b = v
            delta = a_pos_2 - b_pos_2
        else:
            b = u
            delta = b_pos_1 - a_pos_1

        assert delta > 0, "Move Score is Negative"
        moves = moves + delta

    return moves/a.size
```

Bibliography

[1] Food {and} Drug Administration (FDA). *Understanding Barriers to Medical Device Quality*. Oct. 31, 2011. URL: https://www.fda.gov/media/82284/download (visited on 04/14/2022).

[2] *1.1. Linear Models*. scikit-learn. URL: https://scikit-learn/stable/modules/linear_model.html (visited on 04/12/2022).

[3] *1.11. Ensemble methods*. scikit-learn. URL: https://scikit-learn/stable/modules/ensemble.html (visited on 04/12/2022).

[4] *1.4. Support Vector Machines*. scikit-learn. URL: https://scikit-learn/stable/modules/svm.html (visited on 04/12/2022).

[5] *2.3. Clustering*. scikit-learn. URL: https://scikit-learn/stable/modules/clustering.html (visited on 04/12/2022).

[6] *6.2. Feature extraction*. scikit-learn. URL: https://scikit-learn/stable/modules/feature_extraction.html (visited on 04/12/2022).

[7] *About Sparta Systems - TrackWise and TrackWise Digital QMS*. Sparta Systems. URL: https://www.spartasystems.com/about-us/ (visited on 04/14/2022).

[8] *Airbus publishes 2018 aircraft list prices*. Business Traveller. URL: https://www.businesstraveller.com/business-travel/2018/01/16/airbus-publishes-2018-aircraft-list-prices/ (visited on 02/10/2022).

[9] Debby Arkell. *Built-In Quality*. Boeing Frontiers Online. URL: https://www.boeing.com/news/frontiers/archive/2003/july/i_ca1.html (visited on 03/29/2022).

[10] Henning Baars and Hans-George Kemper. "Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework". In: *Information Systems Management* 25.2 (Mar. 28, 2008). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10580530801941058, pp. 132–148. ISSN: 1058-0530. DOI: 10.1080/10580530801941058. URL: https://doi.org/10.1080/10580530801941058 (visited on 03/27/2022).

[11] *Boeing celebrates its 10,000th 737 aircraft with a new record*. Guinness World Records. Mar. 20, 2018. URL: https://www.guinnessworldrecords.com/news/commercial/2018/3/boeing-celebrates-its-10-000th-737-aircraft-with-a-new-record-518888 (visited on 02/07/2022).

[12] *Boeing Company - Investors - Fact Sheets*. URL: https://investors.boeing.com/investors/fact-sheets/default.aspx (visited on 04/05/2022).

[13] *Boeing hit with $17 million fine over 737 production mistakes*. The Seattle Times. Section: Boeing & Aerospace. May 27, 2021. URL: https://www.seattletimes.com/business/boeing-aerospace/boeing-will-pay-faa-at-least-17-million-to-settle-737-production-mistakes/ (visited on 02/10/2022).

[14] Alireza Borjali et al. "Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation". In: *Computers in Biology and Medicine* 129 (Feb. 2021), p. 104140. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2020.104140. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482520304716 (visited on 07/26/2021).

[15] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv:1810.04805 [cs]* (May 24, 2019). version: 2. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 02/28/2022).

[16] BENOIT FARLEY. "Extracting information from free-text aircraft repair notes". In: *Artificial Intelligence for Engineering Design, Analysis and Manufacturing : AI EDAM* 15.4 (Sept. 2001). Num Pages: 11 Place: Cambridge, United Kingdom Publisher: Cambridge University Press, pp. 295–305. ISSN: 08900604. URL: http://www.proquest.com/docview/195891182/abstract/516828437A4E4BB6PQ/1 (visited on 07/26/2021).

[17] Ted Fuhr, Evgeniya Makarova, and Vanya Telpis. "Capturing the value of good quality in medical devices | McKinsey". In: (Feb. 24, 2017). URL: https://www.mckinsey.com/industries/life-sciences/our-insights/capturing-the-value-of-good-quality-in-medical-devices#0 (visited on 04/14/2022).

[18] Yvonne Gibbs. *Shuttle Carrier Aircraft*. NASA. Publisher: Brian Dunbar. Aug. 11, 2015. URL: http://www.nasa.gov/centers/armstrong/news/FactSheets/FS-013-DFRC.html (visited on 02/07/2022).

[19] James Jardine. *How to Shift From Proactive to Predictive Quality Data Management*. MasterControl. URL: https://www.mastercontrol.com/gxp-lifeline/how-to-shift-from-proactive-to-predictive-quality-data-management (visited on 04/14/2022).

[20] Gretchen Jezerc. *Quality More Important to Consumers than Price as Influence of Discounts on Purchase Decisions Declines*. URL: https://www.firstinsight.com/press-releases/quality-more-important-than-price-study (visited on 03/29/2022).

[21] Leslie Josephs. *FAA fines Boeing $6.6 million over compliance and quality-control lapses*. CNBC. Section: Airlines. Feb. 25, 2021. URL: https://www.cnbc.com/2021/02/25/boeing-fined-by-faa-over-dreamliner-production-lapses.html (visited on 02/10/2022).

[22] A. Kao, N.B. Niraula, and D. Whyatt. "Part Name Normalization". In: *2019 IEEE International Conference on Prognostics and Health Management (ICPHM), 17-20 June 2019*. 2019 IEEE International Conference on Prognostics and Health Management (ICPHM). Piscataway, NJ, USA: IEEE, 2019, 6 pp. DOI: 10.1109/ICPHM.2019.8819386.

[23] Qi Zheng Lee et al. "Unlocking the True Value of Permanent Acoustic Sensors via Integration in a Digital Field as a Proactive Method of Sand Monitoring in Gas Wells". In: International Petroleum Technology Conference. OnePetro, Mar. 16, 2021. DOI: 10.2523/IPTC-21851-MS. URL: https://onepetro.org/IPTCONF/proceedings/21IPTC/1-21IPTC/D012S045R134/460775 (visited on 04/14/2022).

[24] John B Maggiore. "Remote Management of Real-Time Airplane Data". In: *Aero Quarterly* 2017.3 (), p. 6. URL: https://www.boeing.com/commercial/aeromagazine/articles/qtr_3_07/AERO_Q307_article4.pdf.

[25] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering". In: *The Journal of Open Source Software* 2.11 (Mar. 2017). Publisher: The Open Journal. DOI: 10.21105/joss.00205. URL: https://doi.org/10.21105%2Fjoss.00205.

[26] Nobal B. Niraula, Anne Kao, and Daniel Whyatt. "Part and condition extraction from aircraft maintenance records". In: *2020 IEEE International Conference on Prognostics and Health Management, ICPHM 2020, June 8, 2020 - June 10, 2020*. Vol. 2020-June. Proceedings of the Annual Conference of the Prognostics and Health Management Society, PHM. Detroit, MI, United states: Prognostics and Health Management Society, 2020. DOI: 10.1109/ICPHM49022.2020.9187064.

[27] International Civil Aviation Organization. *ICAO Safety Management Manual Doc 9859*. 2018.

[28] *Pretrained Models — Sentence-Transformers documentation*. URL: https://www.sbert.net/docs/pretrained_models.html (visited on 04/12/2022).

[29] Leslie Proudfit. *SOFIA Overview*. NASA. Apr. 21, 2015. URL: http://www.nasa.gov/mission_pages/SOFIA/overview/index.html (visited on 02/07/2022).

[30] Monisha Pushpanathan. "Inferring insulin regimen from clinical notes : using natural language processing techniques to extract data from free text records". Accepted: 2021-10-08T16:59:27Z. Thesis. Massachusetts Institute of Technology, 2020. URL: https://dspace.mit.edu/handle/1721.1/132855 (visited on 03/26/2022).

[31] Vikas Raunak, Vivek Gupta, and Florian Metze. "Effective Dimensionality Reduction for Word Embeddings". In: *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 235–243. DOI: 10.18653/v1/W19-4328. URL: https://aclanthology.org/W19-4328 (visited on 01/29/2022).

[32] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *arXiv:1908.10084 [cs]* (Aug. 27, 2019). arXiv: 1908.10084. URL: http://arxiv.org/abs/1908.10084 (visited on 02/27/2022).

[33] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. "Parametric UMAP Embeddings for Representation and Semisupervised Learning". In: *Neural Computation* 33.11 (2021). Publisher: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . ., pp. 2881–2907.

[34] *scipy.stats.kendalltau — SciPy v1.8.0 Manual*. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html (visited on 02/19/2022).

[35] Maria Seale et al. "Approaches for Using Machine Learning Algorithms with Large Label Sets for Rotorcraft Maintenance". In: *2019 IEEE Aerospace Conference*. 2019 IEEE Aerospace Conference. ISSN: 1095-323X. Mar. 2019, pp. 1–8. DOI: 10.1109/AERO.2019.8742027.

[36] Abhishek Sharma, Praprut Songchitruksa, and Rajeev Ranjan Sinha. "Integrating Domain Knowledge with Machine Learning to Optimize Electrical Submersible Pump Performance". In: SPE Canadian Energy Technology Conference. OnePetro, Mar. 11, 2022. DOI: 10.2118/208972-MS. URL: https://onepetro.org/specet/proceedings/22CET/2-22CET/D021S021R003/482811 (visited on 04/14/2022).

[37] David Shepardson and Eric M. Johnson. "FAA says new Boeing production problem found in undelivered 787 Dreamliners". In: *Reuters* (July 13, 2021). URL: https://www.reuters.com/business/aerospace-defense/faa-says-new-boeing-production-problem-found-undelivered-787-dreamliners-2021-07-13/ (visited on 02/10/2022).

[38] Steven Spear and H. Kent Bowen. "Decoding the DNA of the Toyota Production System". In: *Harvard Business Review* 77.5 (Oct. 9, 1999). Publisher: Harvard Business School Publication Corp., pp. 96–106. ISSN: 00178012. URL: https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=2216294&site=ehost-live&scope=site&authtype=shib&custid=s8978330 (visited on 03/29/2022).

[39] Sparta Systems. *QualityWise.ai Auto-Categorization*. URL: https://go.spartasystems.com/rs/084-QBA-512/images/Datasheet-QualityWise.ai_Autocategorization.pdf (visited on 04/14/2022).

[40] Hirotaka Takeuchi and John Quelch. "Quality Is More Than Making a Good Product". In: *Harvard Business Review* (July 1, 1983). Section: Customer service. ISSN: 0017-8012. URL: https://hbr.org/1983/07/quality-is-more-than-making-a-good-product (visited on 03/29/2022).

[41] Andrea Trujillo, Marcos Orellana, and María Inés Acosta. "Design of Emergency Call Record Support System Applying Natural Language Processing Techniques". In: *Information and Communication Technologies of Ecuador (TIC.EC)*. Ed. by Efrain Fosenca C et al. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020, pp. 53–65. ISBN: 978-3-030-35740-5. DOI: 10.1007/978-3-030-35740-5_4.

[42] *U.S. light vehicle market share by automotive manufacturer 2019*. Statista. URL: https://www.statista.com/statistics/249375/us-market-share-of-selected-automobile-manufacturers/ (visited on 03/29/2022).

[43] Michael Wayland. *Toyota's per-car profits lap Detroit's Big 3 automakers*. The Detroit News. URL: https://www.detroitnews.com/story/business/autos/2015/02/22/toyota-per-car-profits-beat-ford-gm-chrysler/23852189/ (visited on 03/29/2022).

[44] *What is Industry 4.0 and how does it work? | IBM*. URL: https://www.ibm.com/topics/industry-4-0 (visited on 03/29/2022).

[45] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.