Data-driven approaches to disaster preparedness: Integrating natural language processing and machine learning for enhanced infrastructure system resilience

Julia Carier Lensing

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

John Y. Choe, Chair

Marc O. Eberhard

Jeffrey W. Berman

Prashanth Rajivan

Program Authorized to Offer Degree:
Industrial and Systems Engineering

University of Washington

Abstract

Data-driven approaches to disaster preparedness:
Integrating natural language processing and machine learning
for enhanced infrastructure system resilience

Julia Carier Lensing

Chair of the Supervisory Committee:
John Y. Choe
Department of Industrial and Systems Engineering

In an era marked by intensifying disasters caused by natural hazards, improving resilience and preparedness has become increasingly important for safeguarding communities and the critical infrastructures on which they depend. Earthquakes, in particular, pose severe threats to human well-being and the integrity of critical infrastructure systems, such as bridges, which serve as key conduits for transportation, emergency response, and economic continuity during and after crises. The rapid acceleration of data generation presents opportunities and challenges; if effectively leveraged, diverse and complex data sources can support more robust, data-driven decision-making for disaster preparedness. This dissertation presents three analytically rigorous and adaptable frameworks to address key challenges using high-dimensional, real-world data to improve disaster preparedness and infrastructure assessment.

This dissertation first presents a text mining framework and an accompanying open-source code designed to extract insights from a novel corpus of practical disaster reports. This chapter highlights the ability to synthesize lessons from past disasters, providing observations that inform future preparedness initiatives. Second, this work introduces a machine learning-based framework for predicting key bridge characteristics related to seismic vulnerability. This framework reduces the need for manual data collection while increasing the availability of network-level characteristics,

thus allowing for a more comprehensive understanding of infrastructure resilience. Third, this work proposes an integrated framework that combines natural language processing, feature engineering, and uncertainty quantification to improve the reliability and interpretability of seismic vulnerability assessments based on small, information-rich datasets.

Together, these contributions lay the groundwork for future research and practice in disaster resilience, particularly in contexts characterized by limited data availability and high data complexity. By demonstrating how natural language processing and machine learning can be used to harness textual, structured, and high-dimensional data, we aim to advance intelligent, data-efficient methods that support more informed and effective decision-making in disaster preparedness and infrastructure system management.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, John, for the countless hours you devoted to providing insight, encouragement, and thoughtful guidance throughout this journey. Your patience, steady presence, and unwavering support helped me find clarity whenever I felt lost, and I am deeply grateful for your commitment to my growth as both a scholar and a researcher.

To Marc and Jeff—thank you for teaching me more about bridges than I ever imagined I'd need to know. Your deep expertise and generous willingness to share your time and knowledge were invaluable to this work.

Thank you to Scott for always being willing to serve as an extra set of eyes on a paper and a constant source of encouragement, not only during this dissertation but throughout my academic and professional pursuits.

I am also sincerely grateful to my committee members, professors, mentors, and colleagues, whose support, encouragement, and belief in me—especially during the most challenging moments—have shaped this dissertation and helped me complete it.

DEDICATION

To my husband, Matt — For your unwavering belief in me, heartfelt encouragement, and quiet but powerful reminders that I belonged, even when I doubted myself.

To Maddie, Brooke, and Liam — You didn't need to understand the dissertation to be its greatest supporter. You fueled it in other ways—with food, chaos, patience, and the kind of love that kept me going even on the hardest days.

Chapter 1

INTRODUCTION

In an era marked by intensifying disasters caused by natural hazards [78], improving resilience and preparedness has become increasingly crucial for safeguarding communities and the critical infrastructures on which they depend. Earthquakes, in particular, pose serious threats to human well-being and the integrity of essential structures, such as bridges, which serve as vital conduits for transportation, emergency response, and economic continuity during and after crises. Simultaneously, data generation has accelerated to unprecedented levels [104], providing vast resources that, if effectively leveraged, could enhance disaster preparedness and resilience efforts. The volume of data has increased, and with it, the availability of advanced analytical tools offers new possibilities to extract valuable insights and make data-driven decisions to prioritize infrastructure system safety and reliability.

This dissertation explores the application of modern data science techniques to address key gaps in disaster science and infrastructure risk assessment by harnessing underutilized data sources and developing scalable, interpretable, and data-efficient frameworks. These adaptive frameworks allow them to accommodate evolving data environments, diverse risk contexts, and changing infrastructure needs. The three chapters that follow each target a specific challenge—from synthesizing unstructured disaster reports to predicting seismic vulnerabilities of transportation infrastructure—and offer innovative methodological solutions to improve the reliability, accessibility, and actionable value of disaster-related information.

Chapter 2 addresses the often overlooked potential in analyzing unstructured textual data in practical disaster reports. While rich in insight, these documents are typically lengthy, dense, and highly technical — characteristics that limit their accessibility and usefulness to a broader audience

and hinder the extraction of actionable insights[35]. Moreover, agencies often tailor these reports narrowly to specific events, which reduces their broader applicability [35], and they rarely store them in structured repositories, contributing to a persistent knowledge gap between research and practice [3]. Practitioner perspectives—essential to successful disaster response and recovery—remain underrepresented in academic literature[106], while fragmented content across strategic, operational, and tactical levels further complicates their practical application[12]. To address these issues, we introduce a text mining framework, supported by open-source code, to synthesize a novel corpus of practical disaster reports. This framework facilitates the extraction of actionable insights, bridging the gap between strategic lessons and tactical implementation. In doing so, it advances disaster preparedness by increasing the accessibility of practitioner-derived insights and facilitating their integration into planning and policy efforts.

Chapter 3 tackles the scarcity of reliable data needed to assess the seismic vulnerability of transportation infrastructure at a network scale. Traditional experimental and field data collection approaches are constrained by time and expense, making them impractical for large-scale vulnerability assessments. Moreover, high-quality data for such assessments is often limited due to the resource-intensive nature of experimental testing and simulations, especially for less common but critical failure modes such as shear-critical behavior[64]. While simulation-based datasets offer physics-informed, validated outputs, they remain underutilized due to cost and implementation barriers[27]. Researchers increasingly recognize the potential of machine learning to address these gaps, yet its application to infrastructure datasets remains relatively underexplored [108]. This chapter presents a machine learning framework for predicting key bridge characteristics relevant to seismic vulnerability, leveraging data from the National Bridge Inventory (NBI). By automating this process, the framework enables more comprehensive, scalable assessments of bridge networks, significantly reducing data collection burdens while expanding analytical capability.

Chapter 4 focuses on maximizing the utility of limited but information-rich datasets. The challenge of handling large, complex feature sets with limited sample sizes can lead to overfitting and hinder the model's generalizability [80]. Furthermore, existing methods often overlook the uncertainty in the predictions, reducing the reliability and interpretability of the model results,

particularly when working with small data sets [20, 114]. This chapter introduces an integrated framework that combines feature engineering using natural language processing (NLP) of categorical variables with uncertainty quantification (UQ) methods to address these challenges. The result is a robust and data-efficient approach for predicting bridge characteristics and the corresponding *Shear-Critical Column Index (SCCI)*, complete with prediction intervals that provide insight into model reliability.

Chapter 2

TEXT MINING OF PRACTICAL DISASTER REPORTS: CASE STUDY ON CASCADIA EARTHQUAKE PREPAREDNESS

This chapter presents research published in *PLOS One* and can be accessed at `https://doi.org/10.1371/journal.pone.0313259`

2.1 Abstract

Many practical disaster reports are published daily worldwide in various forms, including after-action reports, response plans, impact assessments, and resiliency plans. These reports serve as vital resources, allowing future generations to learn from past events and better mitigate and prepare for future disasters. However, this extensive practical literature often has limited impact on research and practice due to challenges in synthesizing and analyzing the reports. In this study, we 1) present a corpus of practical reports for text mining and 2) introduce an approach to extract insights from the corpus using select text mining tools. We validate the approach through a case study examining practical reports on the preparedness of the U.S. Pacific Northwest for a magnitude 9 Cascadia Subduction Zone earthquake, which can potentially disrupt lifeline infrastructures for months. We conducted a brief survey of potential user groups to explore opportunities and challenges associated with text mining of practical disaster reports. The case study illustrates the types of insights that our approach can extract from a corpus. Notably, it reveals potential differences in priorities between Washington and Oregon state-level emergency management, uncovers latent sentiments expressed within the reports, and identifies inconsistent vocabulary across the field. Survey results highlight that while simple tools may yield insights primarily interpretable by experienced professionals, more advanced tools utilizing large language models, such as Generative Pre-trained Transformer (GPT), offer more accessible insights, albeit with known risks associated with current artificial intelligence technologies. To ensure reproducibility, all supporting data and

code are made publicly available (DOI: 10.17603/ds2-9s7w-9694).

2.2 Introduction

The Cascadia Subduction Zone (CSZ) that stretches 700 miles from Vancouver Island, Canada to Northern California is capable of producing up to a 9.0 magnitude (M9) megathrust earthquake [25] that would cause expansive damage and disruption across southern British Columbia, Washington, Oregon and north California [79]. As the interval since the last M9 earthquake in 1700 exceeds the average interval between CSZ events, local, state, regional, and federal emergency management professionals have conducted planning conferences and emergency response exercises to increase awareness of potential impacts and build preparedness for an M9 Cascadia megathrust earthquake. Emergency management practitioners capture critical lessons learned, actions taken, and planning considerations from these practical events in after-action reports, response plans, impact assessments, resiliency plans, and other practical reports. These reports are then shared across the field to inform future preparedness events and expand the knowledge base.

These documents are often lengthy and packed with dense information across a wide range of topics and fields, making them difficult for a reader to digest. They are also usually geared towards a specific disaster event, locality, or infrastructure, which can discourage readers from reviewing them when it does not directly relate to their area of interest. Though these reports contain a wealth of information, individually, they likely have limited readership due to such issues. We propose examining these practical reports as a corpus (body of work as a whole) to increase their accessibility and applicability to a broader audience. Examining a corpus may uncover common trends and themes amongst the vast literature, distinctive features that set one document apart from another, emerging patterns throughout a time period, relationships and connections between documents or represented entities, or underlying emotions across the corpus. Examining a corpus as an aggregate of information-rich individual reports may reveal insights not visible from examining single documents.

To do this, the reader must be able to efficiently cull through hundreds of pages of text and systematically extract and understand important information. Reading and synthesizing even one

of these reports is a lengthy and cumbersome process; conducting manual contextual analysis of several reports at once can be out of reach for busy researchers and practitioners. Text mining using statistical software can effectively and efficiently extract statistical and linguistic features from a collection of documents to look for the presence or absence of patterns, trends, associations, distributions, and sentiments [34]. This paper presents a suite of text mining tools that captures the breadth of existing capabilities while remaining easy to implement, even for users with minimal coding experience.

We provide a flexible script of all the presented tools that can be tailored and applied to various document collections. Researchers who may grasp domain-related theory but have little or no coding experience could use such tools to define and understand the research-practice gap by examining what does and does not appear in such application-oriented corpora. Practitioners may also benefit from the suite of tools to gain objective insights into a corpus, thus complementing their traditional research and other information-gathering approaches (e.g., literature review, environmental scan, landscape analysis). Our suite of tools can be customized to the practitioner's preferences, serving as a valuable time-saving tool to gain high-level, productive comprehension of the large amounts of data in various corpora.

The main contributions of this work are 1) the presentation of a novel practical report corpus that has not previously been explored using text mining, 2) an approach to extract insights from the corpus using select text mining tools, and 3) a flexible script that researchers and practitioners can apply to a variety of corpora. We validate our approach by implementing our suite of tools on our case study corpus and conducting a brief survey of emergency management professionals to elicit feedback on our proposed approach. Lastly, we identify key insights from the individual tools presented and synthesize our findings to present practical implications for emergency management and other risk professionals, including researcher and practitioner user groups.

The remainder of the paper is organized as follows: Background provides an overview of text mining procedures and applications. Approach details the procedures for developing our corpus, and describes the text mining tools' use in our approach. In Case study: Application, results, and interpretations, we present a case study of the application of the approach along with results and

their interpretations, and findings from our informal survey. Finally, we conclude in Discussion with strengths and limitations of our proposed approach, implications for our identified user groups, and recommendations for future work.

2.3 Background

2.3.1 Foundations and modern advances in text mining

In broad terms, text mining is a "knowledge-intensive process in which a user interacts with a document collection over time using a suite of analysis tools" [34]. The available tools, techniques, algorithms, and applications of text mining continue to grow in size and scope as the development and availability of web-based textual information continues to grow [29]. Text mining involves techniques from natural language processing (NLP), machine learning, statistical analysis, and artificial intelligence (AI) to extract meaningful information from dense, unstructured textual data. Users apply it across various domains and applications, including information retrieval, content analysis, social media monitoring, market research, customer feedback analysis, fraud detection, and healthcare analysis. Regardless of the applied tool, text mining involves the same steps. It begins with identifying the document collection. These corpora vary widely in size, content, and format, and may cover a broad range of topics or focus on a specific domain or subject. The surge of textual information on web-based media extends to other sources, including websites, emails, and social media posts [29]. The next step is to convert textual, unstructured data from the corpus into structured, machine-readable elements. This involves preprocessing operations such as removing irrelevant information for the task at hand (removing numbers or 'stop words'), grouping different inflected forms of the same word (lemmatization), or correcting any grammatical or contextual errors (misspellings, abbreviations, case correcting, etc.) [46]. Next, the remaining textual information in each document is converted into a numerical representation. The most commonly used representations are Bag-of-Words (BoW), tf-idf (combination of the word's frequency in a document with its rarity across all documents) [56] and, more recently, *word embedding* [5]. Various analytical tools can be applied once the textual data is preprocessed and converted into a suitable format. These include categorization and classification (assigning documents to predefined categories or

labels) [58], clustering (grouping similar documents) [109], topic modeling (identifying latent topics within the text), sentiment analysis (understanding the sentiment or opinion expressed in the text), summarization (creating a short, coherent representation of a larger document) [41], among many others. Finally, the results are evaluated for quality and relevance and then synthesized with background knowledge/expertise to develop actionable insights.

In this work, we follow these established procedures, adhering to accepted practices in the field to guide our analysis. However, as the tools for text mining evolve, particularly with the integration of advanced artificial intelligence (AI) models, the ability to extract more profound, more comprehensive insights from unstructured textual data has become increasingly more robust. Recent advancements in Large Language Models (LLMs), Deep Learning, and Natural Language Processing(NLP) further enhanced traditional text mining, offering new approaches to real-time data analysis and text interpretation, particularly relevant in disaster management contexts. For example, developing LLM-assisted platforms enables faster, more efficient crisis management, allowing real-time collaboration and decision-making during emergencies[82]. In parallel, deep learning methods for tweet classification have been applied to prioritize and schedule rescue efforts by classifying critical information from social media posts during Hurricanes Harvey and Irma [54]. Additionally, hybrid and ensemble models in NLP have further improved the accuracy of everyday text mining tasks such as classification, clustering, and sentiment analysis, making these tools increasingly valuable for analyzing disaster reports [51].

While these advancements demonstrate the potential of modern AI techniques, text mining remains a foundational approach that effectively complements these new tools. By utilizing established text mining tools, we can systematically extract valuable insights from large volumes of unstructured data, providing a robust framework for analysis. Text mining is particularly well-suited for exploring specific nuances and contextual details found in disaster-prone reports, which more generalized AI models may overlook. Therefore, rather than making traditional text mining obsolete, these advancements enhance its effectiveness, enabling a more in-depth and detailed understanding of the text. Integrating established methods with innovative technologies can equip researchers and practitioners with the tools needed to address the complexities inherent in disaster

management.

2.3.2 Motivation for exploring disaster preparedness practical reports

This study presents a corpus of practical reports related to emergency preparedness for a Cascadia M9 megathrust earthquake. It is essential to analyze this corpus because it represents the unique perspective of emergency management practitioners and can aid in bridging the knowledge gap between research and practice. Albris et al. [3] identifies "a need for synthesis and compilation of lessons learned." Doing so can lead to innovation and change across the emergency management community. Specifically, for public health emergency preparedness and response (PHEPR), National Academies of Sciences, Engineering, and Medicine [75] calls for "a national repository of [after-action reports (AARs)] or of reports based on analysis of AARs" for dissemination of lessons learned to support PHEPR practices and demonstrate a labor-intensive method to rigorously synthesize research studies and reports (e.g., AARs, case reports). Additionally, a body of knowledge from the practitioner's perspective represents factors that can be difficult to convey in research (e.g., experience, relationships, etc.) and are significant determinants of success during disaster recovery [106]. Finally, Boin and 't Hart [12] distinguishes between an emergency response system's strategic, operational, and tactical levels and identifies their unique requirements and contributions to effective disaster recovery practices. Both perspectives are valuable but differ in size, scope, and impact. By extracting valuable insights from the practitioner, we marry these two perspectives to ensure continuity and unity at all echelons [98].

This study presents an effective and efficient method for compiling important insights, knowledge, and lessons learned from practitioners' perspectives.

2.3.3 Relevant corpora that have been explored using text mining

In the emergency management field, text mining is increasingly used to glean information from scientific and academic corpora. Text mining has been used to identify important aspects of resilience in emergency management [74], determine critical success factors in disaster management [17], and analyze resiliency of critical infrastructures [45, 114].

The introduction of social media platforms such as Facebook, X (formerly Twitter), and Instagram precipitated the development of a new category for corpora. Text mining has been used in disaster recovery by extracting locations and failure intensities of infrastructure during the 2015 flood in Chennai, India [19], examining infrastructure-related status [31] and underlying public emotion post-disaster [8] from Twitter data, surveying public sentiment from Tweets following 2018 earthquakes in Lombok and Bali [111], and examining social cohesion of impacted populations following significant hurricanes Harvey, Irma, and Maria [42].

More recently, text mining tools have been applied to other collections of government reports to synthesize general concepts from large textual data sets quickly. Fu et al. [40] applied topic modeling and content analysis of a corpus of urban resilience plans to extract meaningful climate change and urban planning information efficiently. Nam and Nam [73] explored a corpus of COVID-19 reports using feature extraction to understand the global environment surrounding COVID-19 management, and Rubaltelli et al. [92] used text mining to understand emotions surrounding the suspension of a COVID-19 vaccine. Other researchers applied text mining tools to a corpus of annual business reports (10-K) to assess organizational cybersecurity attitudes [50].

Perhaps the most similar to our proposed corpus is that of accident reports. Researchers in the risk analysis field have employed text mining tools to evaluate common cause factors for coal mining accidents [88], pipeline incidents [61, 55], and rail accidents [15].

Our study explores a new corpus of practical reports and demonstrates the ability to extract insights from the practitioner's unique perspective.

2.4 Approach

In this section, we outline the scope of our corpus of practical reports and then discuss the details of our document and feature analysis, sentiment analysis, and topic modeling and summarization. We conclude with our survey protocols.

2.4.1 Corpus development

As discussed above, many different corpora have been used for text mining, but there has been limited exploratory work on practical reports as a corpus. Practical reports are used across various fields (e.g., health sciences, military, engineering, emergency/disaster management) to document procedures, findings, and outcomes of specific experiments, practical activities, or other hands-on exercises. In the emergency management field, common practical reports include after-action reports, resiliency assessments, response plans, and planning scenarios. While each of these reports serves a distinct function, they contain interconnected elements of a comprehensive approach to enhancing emergency preparedness, response, and overall resilience. We use this set of documents and the definitions below to establish the scope of our corpus.

After-action reports (AARs) evaluate a completed activity, mission, or event, capture lessons learned, identify strengths and weaknesses, and recommend future improvement. Resiliency assessments assess an organization's capacity to withstand and recover from adverse situations, challenges, or crises, and measure various factors that contribute to resilience. A response plan documents procedures, actions, and protocols to guide individuals, organizations, or communities in effectively managing and responding to emergencies, crises, disasters, or other unexpected events. A planning scenario is a hypothetical situation or set of circumstances used to develop plans, strategies, or exercises to prepare for and respond to emergencies, crises, or other events.

2.4.2 Suite of tools for text mining analysis

Below, we discuss our suite of text mining tools. Our suite of tools can be applied to any practical report corpus to understand its makeup, represent its sentiment, and identify its common themes and ideas. This approach provides a baseline, exploratory overview of a corpus that serves as a jumping-off point for more detailed follow-on analysis, and can help answer our research question: *What are the unique or generalizable characteristics in this corpus that represent the relevant past efforts in practice?*

Document and corpus feature analysis

In document and corpus feature analysis, we explore the individual terms that form a document. We can examine terms at the document level or aggregate them at the corpus level. The tools shown here help us answer questions like *what are the predominant themes in this corpus?* and *what are the unique or characteristic themes in each document of the corpus?*. We can also use these tools to compare terms between documents.

To conduct feature analysis, we use a Bag-of-Words (BoW) model that transforms each document into a vector such as $v = [x_1, x_2, ..., x_n]$, where x_i represents a term frequency or term frequency-inverse document frequency (tf-idf)[110]. The term frequency is a count of the occurrence of words in a single document. Commonly, high-frequency words represent the document's general concepts and ideas. Tf-idf helps the researcher identify the critical words in each document by placing heavier weight on the words used less frequently and decreasing the weight for words or terms used frequently and commonly throughout the corpus[97]. It is represented by

$$\text{tf-idf}(w, d) = \text{TermFreq}(w, d) \cdot \log\left(\frac{N}{\text{DocFreq}(w)}\right), \tag{2.1}$$

where $\text{TermFreq}(w, d)$ is the frequency of the word w in the document d, N is the total number of documents, and $\text{DocFreq}(w)$ is the number of documents containing the word w [34]. This analysis explores these vector representations as is and in more complex modeling efforts described below.

Sentiment analysis

Using our vector representations from feature analysis, we can also explore the sentiment or feeling across the corpus and answer the question *what is the portrayed sentiment of this document or corpus?*

In this work, we use two general-use lexicons - Bing and NRC. The Bing Liu and collaborators (Bing) lexicon categorizes English words into a binary negative or positive sentiment category. It counts those terms with the most significant contribution to overall sentiment [47]. Similarly, the

National Research Council (NRC) lexicon [70, 72] associates words with two sentiments (negative and positive) but also categorizes terms into eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) [87].

Topic modeling and summarization

Topic modeling assigns a collection of texts into natural groupings to understand the whole collection better, uncover hidden themes from within a collection, assign documents to discovered themes, or use these assignments to organize or summarize the collection [23]. It can aid in answering the questions *what are the most common topics discussed in the corpus?* and *which documents comprise identified topics.* Text summarization involves creating a short, coherent representation of a larger document or corpus by distilling the most important information from a source [41] and can be used to answer the question *how can you summarize this content?*

In this work, we conduct topic modeling using Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representations from Transformers (BERT). Finally, we utilize Generative Pre-trained Transformers (GPT) models to conduct text summarization on identified topics.

Latent Dirichlet Allocation (LDA) Latent Dirichlet Allocation (LDA) is a generative probabilistic model commonly used in topic modeling. It describes each document in the corpus as a collection of topics and each topic as a collection of words. LDA is an iterative process that uses probability distributions to work backward to identify which topics would have generated these documents, and which words would have generated those topics. The researcher determines the number of topics k. The algorithm iterates over each document d and each word w to choose the product of the probability that the topic t is contained in document d and the likelihood that the word w is contained in topic t to determine the most relevant topic for the current word. This process continues until a steady state is achieved, rendering the assignment of words w to k-number of topics based on the probability that a word w belongs to a topic t [11].

Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) Up to this point, we have discussed single-dimensional rep-

resentations of documents for analysis (term frequency, tf-idf, etc.). While these tools are simple to implement and easily accessible, they do not account for the order or the relationship of terms [59]. Recently, state-of-the-art large language models (LLMs) have gained traction for their ability to incorporate semantic analysis using word or sentence embeddings (dense, multidimensional representation of words or sentences in the form of vectors), and their pre-training on large, general-use data sets that can then be applied broadly to a variety of corpora [113]. Current popular large language models such as BERT and GPT use embeddings to represent documents and then pass them through a transformer model that uses attention to track relationships in sequential data (e.g., words or terms in a sentence or sentences in a paragraph) [103].

For this analysis, we explore BERT tools by implementing `BERTopic` on our corpus in `Python`. `BERTopic` converts documents into vector representations (embeddings) using a sentence transformer. It clusters the embeddings using a clustering algorithm and then generates the topic of each cluster based on the cluster-level tf-idf (c-tf-idf) [43]. Much like tf-idf discussed previously, c-tf-idf extracts the most important words in each cluster (c) instead of the entire corpus.

We generate a short summary of each topic using GPT tools based on the bag-of-words topic representation produced in `BERTopic`. Unlike BERT, GPT is an autoregressive model that predicts a term based only on the information to the left of the term [89].

2.4.3 Evaluation of suite of tools

To support our claim that the proposed suite of tools can be helpful for researchers and practitioners, we summarize the case study's findings in a 9-page white paper, which takes less than 10 minutes to read and can be found in Appendix A. We shared it with researchers (including social scientists and engineers) and practitioners (including emergency managers, consultants, engineers, coordinators, and advisors), with members of both groups having little to extensive knowledge about the M9 event. We asked these researchers and practitioners, recruited through the authors' professional contacts, to provide anonymous comments through a Google Form on 1) the specific aspects of the tools that they found particularly useful and 2) challenges or limitations they could identify. We offered $35 gift cards as participant incentives. This survey was exempt from institutional review

board (IRB) review.

2.5 Case study: Application, results, and interpretations

In this section, we apply our suite of tools to a developed corpus of practical reports, present results and visualizations, offer possible interpretations, and discuss our survey findings.

The presented results below are specific to this case study's corpus. The results' interpretations are examples of what researchers and practitioners can gain from applying this approach without aiming to be comprehensive.

Due to space limitations, we describe the most significant findings in the section below. Interested readers can find additional insights in Appendix A.

2.5.1 Corpus development

The motivation for this exploratory work sprang from sharing a collection of emergency preparedness reports from social science researchers who studied the lifeline infrastructure recovery that will follow the Cascadia Subduction Zone megathrust earthquake [90]. Valuing the merit of this collection, our team sought to better understand its composition, scope, and distinctiveness through text mining.

Our initial research sample consisted of 27 publicly available reports about disaster preparedness for Cascadia events in Washington, Oregon, and California, as well as regional and federal-level planning documents. Of the 27 documents, 21 were retained for initial analysis. We removed any documents that did not contain the richness of text required for analysis (e.g., phone directories, report covers, presentations) and any document that did not meet the inclusion and exclusion criteria below.

On inclusion and exclusion criteria, we included general practical reports about disaster preparedness for Cascadia earthquake events as well as specific reports related to post-Cascadia event infrastructure recovery (e.g., water, power, transportation, and telecommunications). We excluded reports related to hospital or healthcare emergency preparedness (unless coupled with another infrastructure), academic research articles, news articles, blog posts, and non-technical articles (e.g.,

information for the public).

To ensure that the initial sample adequately represented publicly available literature, we conducted a Google search of state, regional, and federal preparedness-related websites for any additional reports that would make our corpus more robust. For state-level practical reports, we searched on Washington, Oregon, and California official government sites related to Cascadia event preparedness (e.g., Washington Emergency Management Division, Oregon Department of Emergency Management, and California Governor's Office of Emergency Services). We also examined local or state-affiliated research centers that may produce reports (e.g., Pacific Northwest National Laboratory, Oregon State University Extension Earthquake Preparedness, Cascadia Region Earthquake Workgroup). We manually reviewed identified websites, searched for "Cascadia" on each search bar, and reviewed hits that met the inclusion criteria. To ensure we did not miss any important documents, we also conducted a generalized Google search with relevant keywords ("*state* Cascadia subduction zone disaster preparedness practical reports") to capture state-level documents. A similar process was conducted for regional and federal entities. We reviewed websites for the Federal Emergency Management Administration (FEMA), National Oceanic and Atmospheric Administration (NOAA), Earthquake Engineering Research Institute (EERI), and the United States Geological Survey (USGS), plus a general Google search. Finally, we conducted a Google search with infrastructure-based keywords ("water", "utility", "electricity", "transportation", "telecommunication") to identify any preparedness reports from individual infrastructure entities. This search yielded an additional 17 reports for inclusion in the corpus.

Our corpus is representative of the type of practical reports that are publicly available and intended to be exhaustive concerning the aforementioned inclusion and exclusion criteria at the time of our search from January to March 2022. We searched manually and took relevance, quality, and complexity into account to determine the comprehensiveness of the corpus. We concluded our search when we stopped discovering new information relevant to our objectives. We also took into account maintaining consistency in our search procedures (e.g., depth of search state-to-state and among infrastructures).

Our final corpus comprises 38 practical reports representing Cascadia earthquake preparedness

planning at the state, regional, and federal levels (see Table 2.1 for summary). There are 29 reports from the state level, seven from the regional level (encompassing WA, OR, CA), and two from the federal level. Of the 29 state-level reports, two are from California, 21 from Oregon, and six from Washington. This numerical difference between the states may be attributed to varying interest levels, likely because the longest and most central portion of the Cascadia subduction zone abuts Oregon, followed by Washington and California in the U.S., as shown in Figure 2.1. Given the inclusion criteria and search protocols for our study's purpose, the over-representation of Oregon is considered fair. The equitable representation of state-level reports can and should be considered for the corpus design process if the study's purpose is to compare states. The average page length of the documents in the corpus is approximately 80 pages, with the shortest document at four pages and the longest at 341 pages. The oldest document was published in 1997; the most recent was in 2022.

2.5.2 Document and corpus feature analysis

To conduct document and corpus feature analysis, sentiment analysis, and topic modeling using LDA, we use the `R` programming language and 'tidy' text principles and tools [97]. We conduct pre-processing operations using accepted practices and standardize the representation of each document in our corpus using vectors. Supporting data and code can be found at [60].

In this section, we explore the individual terms that form a document by examining term frequencies and relationships and term frequency-inverse document frequency (tf-idf).

Term frequencies - what are the predominant themes in this corpus?

Figure 2.2 shows the term frequencies across the corpus, sorted in decreasing order, to summarize predominant themes. Some notable trends offer insights into the corpus. For example, Oregon is the second most frequently mentioned, whereas Washington did not make it to this list of the top 15. This trend is likely because 21 out of 29 state reports are from Oregon versus six from Washington. Note that a further investigation below the top 15 may be informative. Still, one should place lesser significance on lower ranks due to Zipf's law (i.e., a term's frequency is approximately inversely

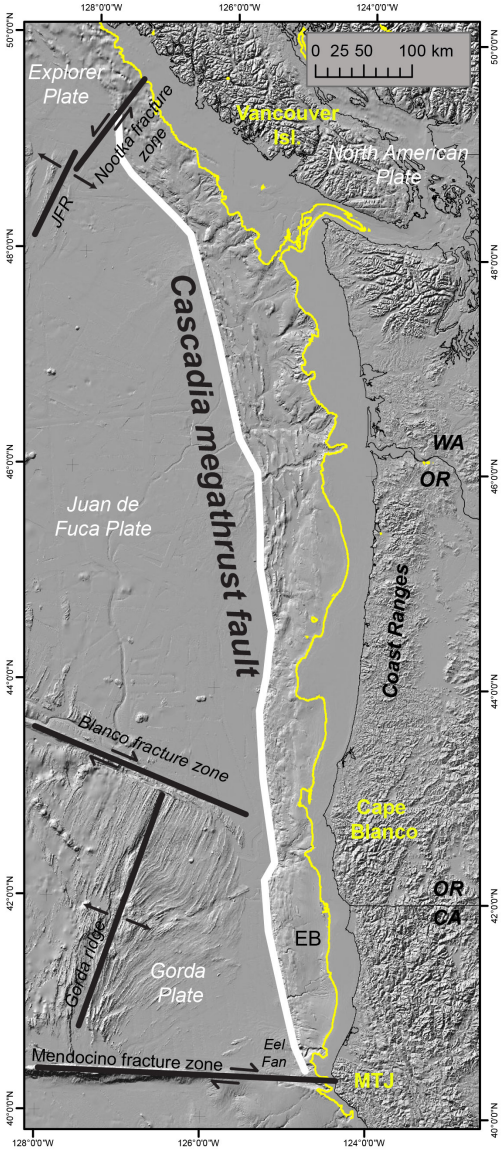


Figure 2.1: Map of the Cascadia Subduction Zone [100]

Document	Page Length	Year	Document Type	Level	State
AAR Cascadia Rising WA State	102	2018	After Action Report	State	WA
AAR Cascadia Rising WA State	38	2022	After Action Report	State	WA
AFIT Basing and Resupply Report	69	2017	Response Plan	Federal	-
Analytical Baseline Study for Cascadia Events	324	2011	Impact Assessment	Federal	-
Blue Cascades After Action Report	66	2018	After Action Report	Regional	-
California CSZ and Tsunami Response Plans	144	2013	Response Plan	State	CA
Cascadia Deep Earthquakes	28	2008	Impact Assessment	Regional	-
Cascadia Earthquake Economics Report	17	2020	Impact Assessment	State	OR
Cascadia Subduction Scenario	30	2013	Planning Scenario	Regional	-
CREW Shallow Earthquakes	32	2009	Impact Assessment	Regional	-
CREW Tabletop Exercises	16	2007	Planning Scenario	Regional	-
CSZ Bridge Impact Report	163	2016	Impact Assessment	State	OR
CSZ Hazard Mapping	147	1997	Impact Assessment	State	OR
Earthquake Scientific Consensus	155	2000	Planning Scenario	State	OR
GHPUD Resilience Study Report	37	2020	Resilience Plan	State	WA
Impacts of CSZ Earthquake on CEI Hub	55	2021	Impact Assessment	State	OR
Joint Multi-State AAR Cascadia Rising	46	2016	After Action Report	Regional	-
Oregon Cascadia PlaybookV3	100	2018	Response Plan	State	OR
Oregon Cascadia Rising AAR	60	2016	After Action Report	State	OR
Oregon DOT Highway Report	19	2013	Impact Assessment	State	OR
Oregon Earthquake Report	157	2012	Impact Assessment	State	OR
Oregon Fuel Action Plan	92	2017	Action Plan	State	OR
Oregon Hospital and Water System Earthquake Risk Evaluation	164	2014	Impact Assessment	State	OR
Oregon Infrastructure Report Card	95	2019	Impact Assessment	State	OR
Oregon Resilience Plan Final	341	2013	Resilience Plan	State	OR
Oregon Transportation Systems Resiliency Assessment	135	2021	Impact Assessment	State	OR
OSSPAC CEI-Hub Report	46	2019	Resilience Plan	State	OR
PWRTF After Action Report Summary	4	2021	After Action Report	State	OR
Resilient Infrastructure Planning Exercise	52	2018	Response Plan	State	OR
Resilient Washington State	40	2012	Resilience Plan	State	WA
Seattle Cascadia Earthquake Workshop Interdependencies	19	2017	Impact Assessment	Regional	-
Vigilant Guard After Action Report	69	2017	After Action Report	State	CA
Washington State Fuel Shortage Action Plan	14	2019	Action Plan	State	WA
Washington State Transportation Resiliency Assessment	69	2019	Resilience Plan	State	WA
Water Interconnection Report	18	2010	Action Plan	State	OR
What to Expect - Climate Change and Earthquakes in Portland	8	2017	Impact Assessment	State	OR

Table 2.1: Corpus summary

proportional to the term's rank).

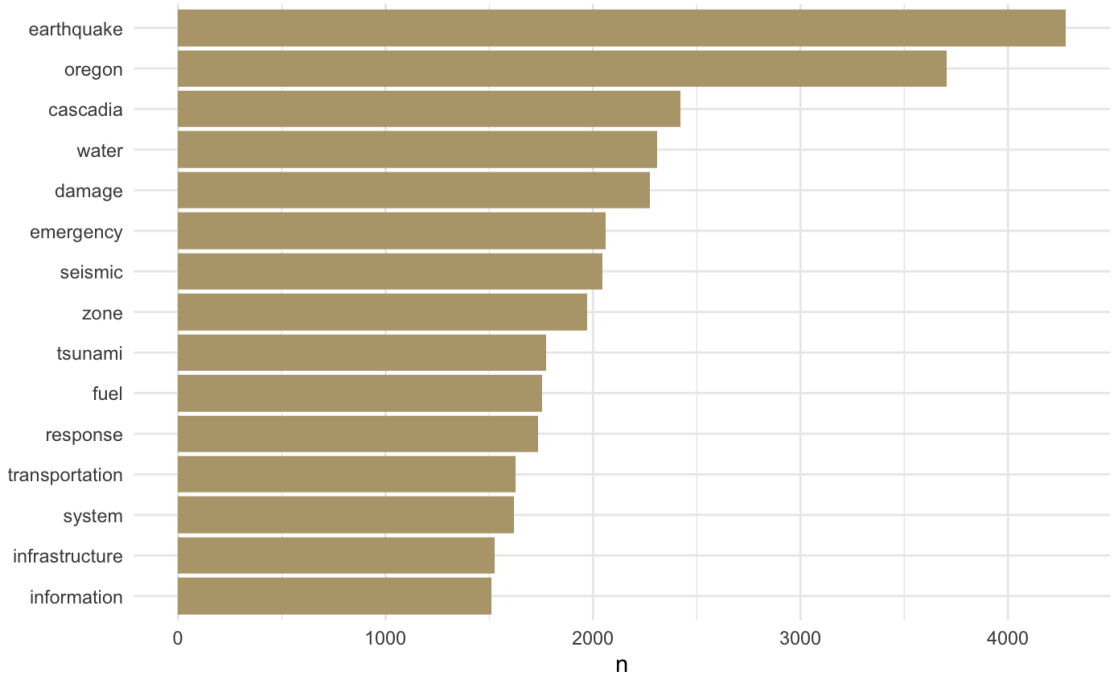


Figure 2.2: Term frequency - top 15 most frequently used single-word terms in the corpus.

Water and transportation are the only two infrastructure systems that reached the top 15. This is surprising because of the commonly perceived importance of the electricity system, upon which other systems' functionality depends. Telecommunications is also notably absent and is widely perceived as a critical infrastructure to facilitate coordination during the immediate disaster recovery period. The higher ranks of *water* and *fuel* over the other infrastructures indicate a greater emphasis on *emergency response* (i.e., the additional two terms appearing in the top 15) over recovery. Therefore, the predominant themes in the corpus seem focused on short-term over long-term actions in the aftermath.

These apparent 'gaps' in the corpus may warrant further investigation to determine their significance and causes. For instance, do these gaps reflect blind spots in practical implementation, where practitioners overlook certain disaster preparedness strategies or information? Alternatively,

do they represent barriers to public access to information, where critical information might not be widely available or easily accessible to relevant stakeholders? Furthermore, these gaps could signal the need to refine the approach by expanding the corpus. This could involve using broader keywords, incorporating more diverse data sources, or tapping into less conventional information repositories to ensure a more comprehensive coverage of disaster preparedness exercises, practices, and insights. Additional exploration may reveal other underlying factors contributing to these gaps, such as the possibility that report writers assume the infrastructure is already sufficiently capable of withstanding significant events, or that certain aspects of infrastructure assessment fall outside their jurisdiction due to a lack of assigned responsibility. The absence of critical infrastructure, like ferry services in Oregon, is a narrow example of the latter, highlighting how infrequent considerations may lead to gaps in preparedness discussions.

Term relationships - what are the predominant themes in this corpus?

Figure 2.3 reveals more in-depth insights than Figure 2.2 into how practical reports use each term in its immediate context. For example, the most frequent terms in Figure 2.2, such as earthquake, Oregon, and Cascadia, commonly appear in specific contexts, e.g., to denote the hazard at issue, namely, Cascadia subduction zone earthquake, or to discuss Oregon-specific preparedness (e.g., Oregon resilience plan, hospital, fuel, transportation systems). By contrast, Figure 2.3 highlights the importance of the following bi-grams in M9 preparedness: private sector, natural gas, mass care, Critical Energy Infrastructure (CEI) Hub, Lincoln City, and Columbia River. For example, the corpus shows practitioners' frequent consideration of the private sector that owns more than 80% of the U.S. critical infrastructure [99]. In sum, these insights could help guide more effective disaster preparedness planning at the local and national levels by identifying critical points of intervention and collaboration.

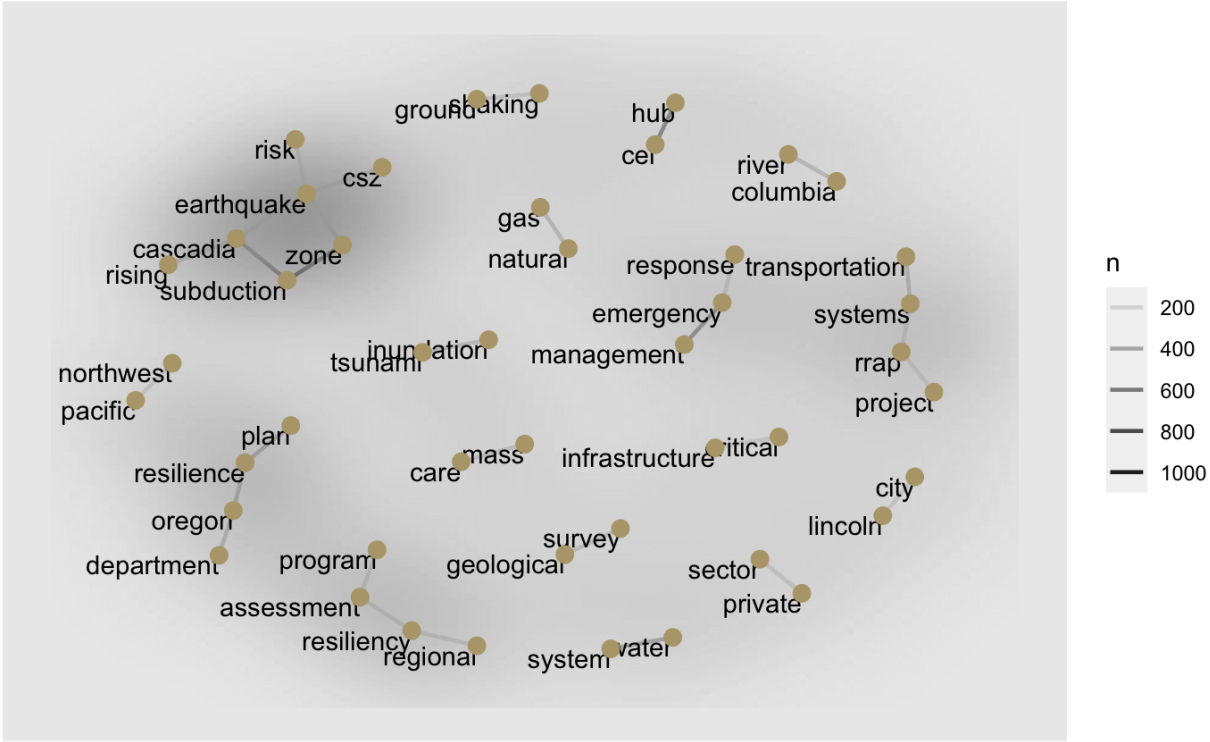


Figure 2.3: Term relationships - bi-grams (2 sequential and adjacent terms) that occur more than 200 times throughout the corpus and their occurrence with other bi-grams/terms. The opacity of the edge and background represents the frequency of the pair (darker = more frequently used, lighter = less frequently used)

Term frequency - inverse document frequency of whole corpus - what are each corpus document's unique/characteristic themes?

Term frequency and inverse document frequency (tf-idf) represent novel concepts, which are particularly emphasized in each practical report. For example, local emergency managers, infrastructure managers, and business continuity practitioners may be surprised to see Figure 2.4, which reveals that only one or two documents pay attention to some terms widely considered critical in practice (e.g., clearinghouse, ESF [Emergency Support Function], hospital, geodatabase, climate, hub, oil, consortium, substation).

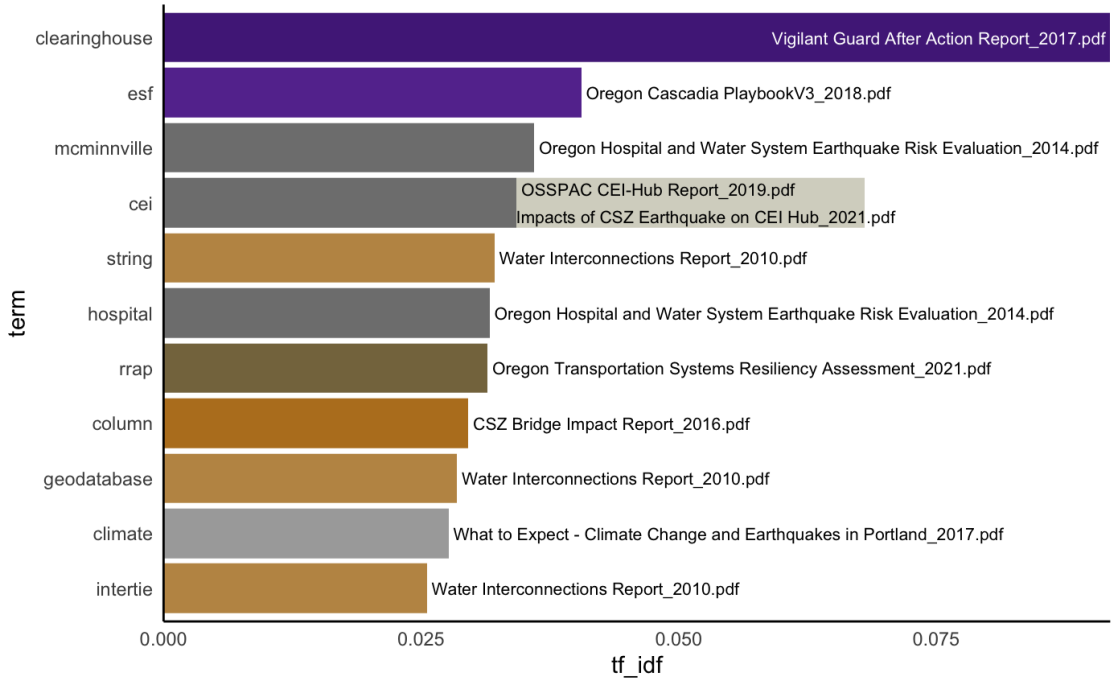


Figure 2.4: Term frequency - inverse document frequency of whole corpus. Notable acronyms are as follows. ESF: Emergency Support Function. CEI: Critical Energy Infrastructure. RRAP: Regional Resiliency Assessment Program.

Another noticeable insight is the use of different terms across various documents, such as clearinghouse, hub, and consortium, all of which emphasize the need for coordination to streamline information sharing and response actions. This analysis echoes the well-documented need for stan-

dardizing the vocabulary and means for communication in emergency management because the lack thereof is widely recognized as a barrier in information sharing [94]. The identified need for an information clearinghouse is one of the immediate takeaways from this analysis of Cascadia practical reports, but its implications extend beyond this specific context. On a broader scale, establishing centralized platforms for information exchange, supported by standardized terminology, is critical for effective disaster response and recovery across various hazard scenarios.

Term frequency - inverse document frequency of select pairs - how do characteristic terms in one document compare to those of another related document?

Figure 2.5 shows characteristic terms in three select pairs of documents. Each term's "characteristicness" within a document is measured by comparing two documents (not the entire corpus), which were identified as similar in their document-topic probabilities (Figure 2.9): 1) Washington state CSZ event exercise AAR in 2018 vs. 2022; 2) state-level resilience planning in OR vs. WA; and 3) state-level transportation systems study through the Regional Resiliency Assessment Program (RRAP) in OR vs. WA.

The first row in Figure 2.5 compares WA AARs on Cascadia Rising (CR) in 2016 (the corresponding report was published in 2017 and revised in 2018) and in 2022. The noticeable shift of characteristic terms (in *italic* in this and the following few paragraphs) from CR16 to CR22. CR16, which unlike CR22 was integrated with the Washington National Guard (*WANG*)'s *Vigilant Guard* exercise, is characterized by the term *guard* and *SAR* (Search and Rescue). This is consistent with operational findings from this report as SAR after the CSZ earthquake will originate not from typically recognized first responders (e.g., national *guard* members), but by ordinary people in the immediate area and community. CR22 emphasizes the *reopening* of critical transportation (e.g., including *surface* transportation) to deliver lifesaving and life-sustaining bulk *goods* and resource support for "impacted, isolated and supporting communities," including *coastal* communities. The COVID-19 pandemic appears to have influenced some of the characteristic terms in CR22 (e.g., *shelter*, *reopening*, *AFN*), which demonstrate the collective knowledge and experience gained through the coordinated response to the unprecedented challenges brought by the pandemic since 2020 (i.e.,

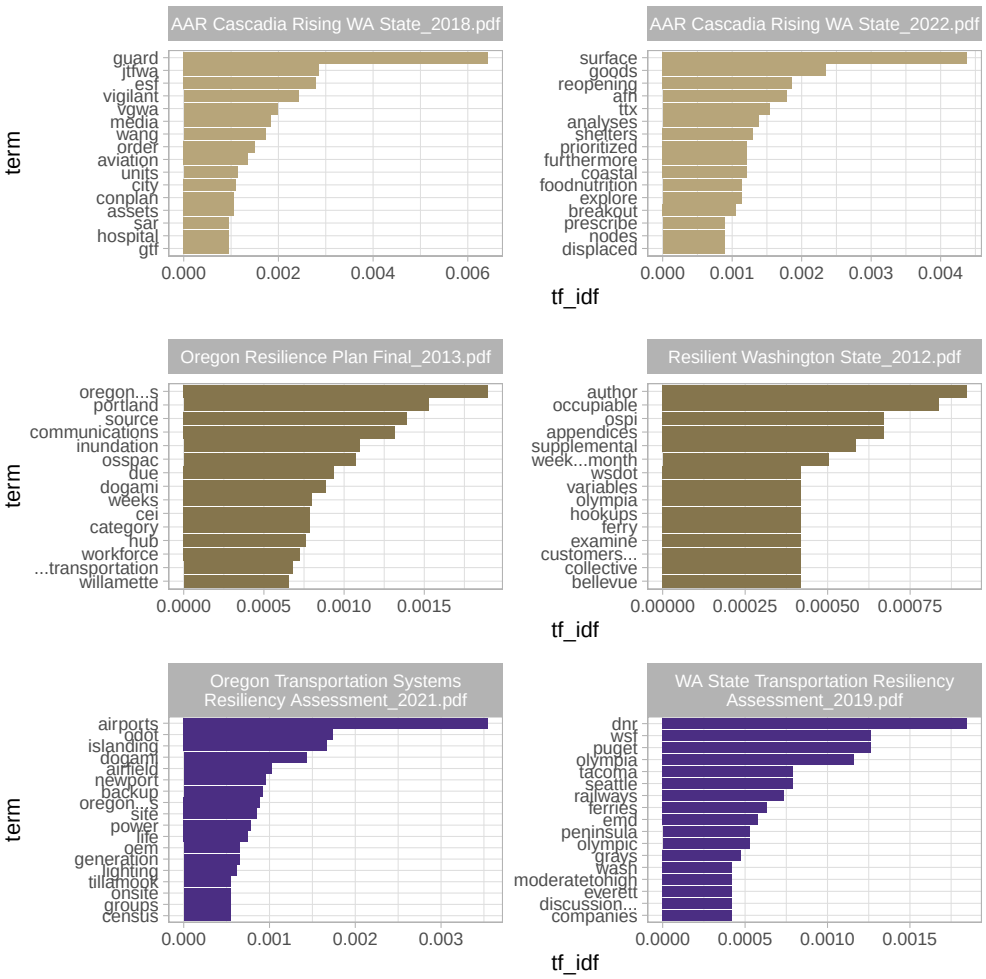


Figure 2.5: Term Frequency - inverse document frequency of select pairs. Notable acronyms are as follows. JTF-WA: Joint Task Force - Washington. VG-WA: Vigilant Guard – Washington. ESF: Emergency Support Function. WANG: Washington National Guard. CONPLAN: Contingency Plan. SAR: Search and Rescue. GTF: Geographic Task Forces. AFN: Access and Functional Needs. TTX: Table-Top Exercise. OSSPAC: Oregon Seismic Safety Policy Advisory Commission. DOGAMI: Oregon Department of Geology and Mineral Industries. CEI Hub: Critical Energy Infrastructure Hub. OSPI: Office of Superintendent of Public Instruction. WSDOT: Washington State Department of Transportation. ODOT: Oregon Department of Transportation. OEM: Office of Emergency Management. DNR: Department of Natural Resources. WSF: Washington State Ferries. EMD: Emergency Management Division. Note: the length differential between Oregon Resilience Plan (341 pages) and Resilient Washington State (40 pages) explains the plateau of tf-idf for WA; a longer document d in Eq. 2.1 tends to have a greater variety of word w which does not appear in the rest of the documents considered in TF-IDF. These unique terms with the same frequency create a plateau.

after CR16).

The second row in Figure 2.5 compares state-level resilience plans: Oregon Resilience Plan (ORP) vs. Resilient Washington State (RWS). tf-idf statistics correctly identify state-specific keywords, such as prominent geographical names (e.g., Portland and Willamette [valley/river] in OR vs. Olympia and Bellevue in WA) and state agencies in charge of the state resilience (e.g., OSSPAC and DOGAMI in OR vs. OSPI and WSDOT in WA). More interesting findings are ORP's emphases on *communications, inundation, workforce*, and *transportation* in contrast to RWS' keywords, such as *occupiable, ferry* and *OSPI*, which is in charge of schools' resilience.

The third row in Figure 2.5 compares state-level transportation resiliency. Like the second row, tf-idf statistics correctly identify state-specific keywords, such as important geographical names (e.g., Newport and Tillamook in OR vs. Puget [Sound], Olympia, Tacoma, Seattle, Olympic Peninsula, Grays [Harbor], and Everett in WA) and state agencies in charge of transportation resiliency (e.g., ODOT, DOGAMI, and OEM in OR vs. DNR, WSF, and EMD in WA). The OR report emphasizes *airports, airfield lighting, islanding*, and *backup power generation* in contrast to the WA report that highlights *railways, ferries*, and *companies*.

The insights discussed in this section provide a broader understanding of how regional differences and temporal factors shape the language and focus of disaster preparedness practical reports. The use of tf-idf can be applied to various domains, enabling researchers and policymakers to identify critical differences, shifts in focus, or areas requiring attention or alignment across different contexts.

2.5.3 Sentiment analysis

Sentiment analysis - what is the portrayed sentiment of this corpus?

A known limitation of sentiment analysis is that it can be challenging to distinguish between topic valence (e.g., the positive or negative sentiment associated with a particular topic) and the author's position on the topic [71]. Figure 2.6 shows the predominantly negative sentiment of the corpus based on the Bing lexicon, which could be representative of the adverse emotions and concerns associated with natural disasters. Conversely, it may prove to be a salient point about the author's choice of words, particularly given the prevalent fatalistic attitude towards the Cascadia

Subduction Zone Earthquake among the public and, more often than not, emergency management practitioners (e.g., local emergency managers and infrastructure managers). Further analysis is required to determine if this finding potentially calls for a paradigm shift in how we present a catastrophic event scenario in practice, to induce a more active and constructive engagement of society in the discussion of catastrophe preparedness. For example, future reports could focus more on positive themes (e.g., what are the benefits of mitigation and preparedness actions?) than negative themes. Future research could study the impact of overall negative sentiment or affect in emergency management practice reports, building upon the relevant risk communication literature that examined the relationship between negative affect and mitigation, preparedness, and protection behaviors [96] [105] [107].

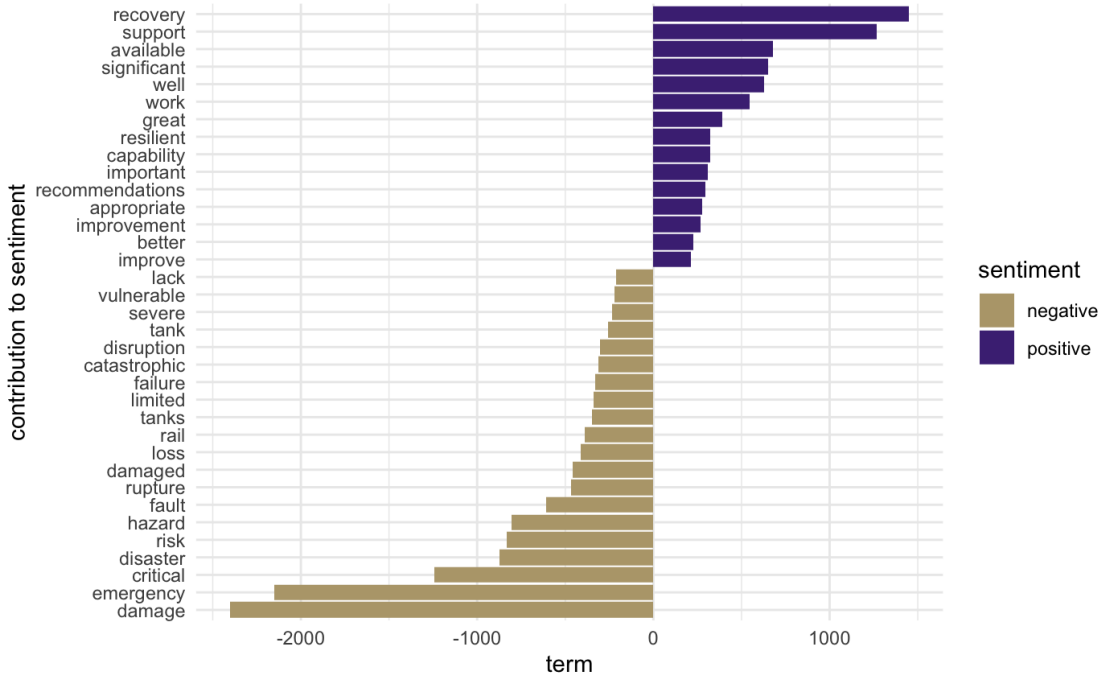


Figure 2.6: Sentiment analysis - count of terms with the greatest contribution to overall sentiment based on BING lexicon.

Sentiment frequency - what emotions are portrayed most in a single document (Oregon Resilience Plan 2013)?

Figure 2.7 offers a closer look at the NRC-based sentiment analysis that maps the document to 8 basic emotions. Most frequent are 'trust'-evoking words, followed by 'fear'-evoking words. This analysis shows the potential value of sentiment analysis for better risk communication, where the impacts of trust and negative affect (e.g., fear, sadness, anger) are extensively studied [105].

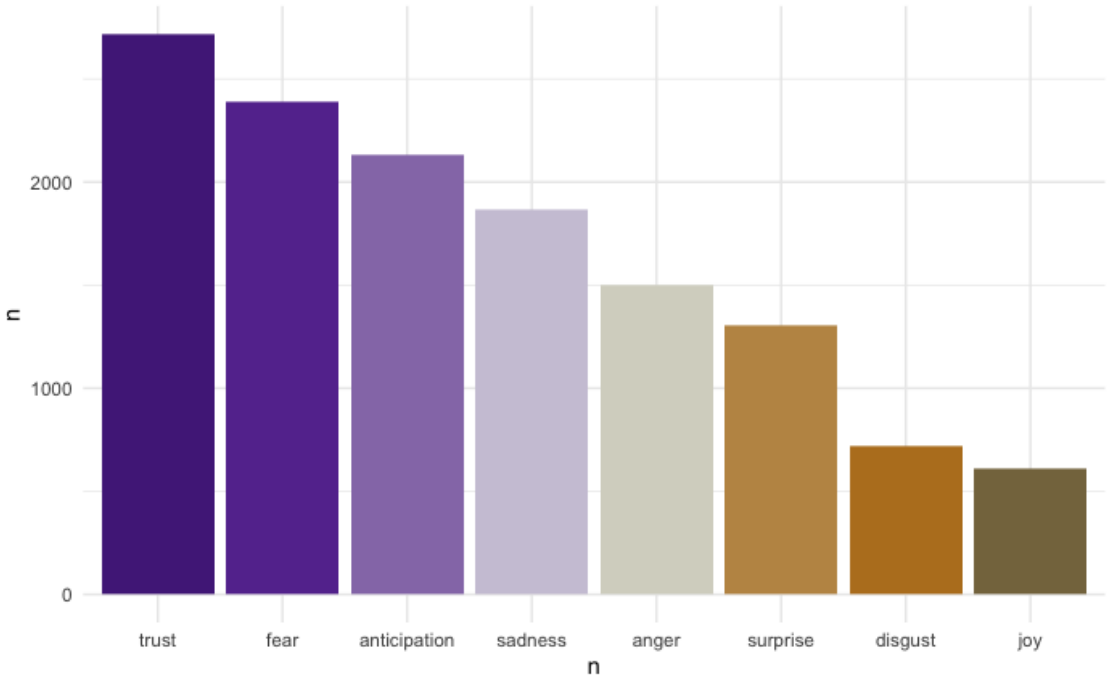


Figure 2.7: Sentiment frequency - frequency of words by sentiment used in Oregon Resilience Plan 2013 according to NRC lexicon. Associates words with eight basic emotions in Plutchik's wheel of emotions. [72].

Sentiment analysis's broader applicability in disaster preparedness reports lies in its ability to provide insight into authors' emotional tone and communication strategies. Identifying the predominant use of negative emotions, such as fear or sadness, could signal the need for a more balanced approach in risk communication. This tool can be applied across various fields where emotional engagement is key, helping to foster more constructive dialogues and encouraging proactive behavior

by balancing negative sentiments with trust and optimism in future reports.

2.5.4 Topic modeling and summarization

Topic modeling with four topics using Latent Dirichlet Allocation (LDA) - what are the most common topics discussed in the corpus?

When using LDA, we need to pre-specify the number of topics (i.e., clusters of terms) to be identified in the corpus. This number may be chosen based on domain knowledge or iteratively tested until reaching a reasonable number of topics (see Topic-word scores using Bidirectional Encoder Representations from Transformers (BERT) - what are the most common topics discussed in this corpus? for `BERTopic`'s ability to select the number algorithmically). For this analysis, we tested several numbers of clusters (see Appendix A for details) and found a 4-topic target to be the most useful for this corpus. A co-author interpreted topic groups, knowledgeable in disaster recovery, who examined commonalities and relationships among identified words to discern topics.

Figure 2.8 presents a subtle breakdown of the topics present in the corpus. Topic 1 highlights the importance of transportation systems in the regional and state (e.g., Washington)-level analysis of earthquake events. Topic 2 describes state-level emergency response plans centered around fuel, information, support, and exercise. Topic 3 identifies the core hazards (i.e., the Cascadia subduction zone earthquake and its induced tsunami), which tend to be illustrated through figures. Topic 4 centers around earthquake damage to water systems and facilities. This last topic's common co-occurrence with 'Oregon' may indicate the perceived primary importance of the water system over other infrastructures to the resilience of Oregon.

Document-topic probabilities using Latent Dirichlet Allocation (LDA) - which documents comprise the identified topics in Figure 2.8?

Figure 2.9 shows how the documents are distributed across the four topics identified in 2.8. Some documents primarily represent a single topic group. For example, the first four documents are characteristic of Topic 1, which emphasizes regional and state transportation systems. This illustration also reveals that a document, 'Cascadia Earthquake Economics Report_2020.pdf', focuses

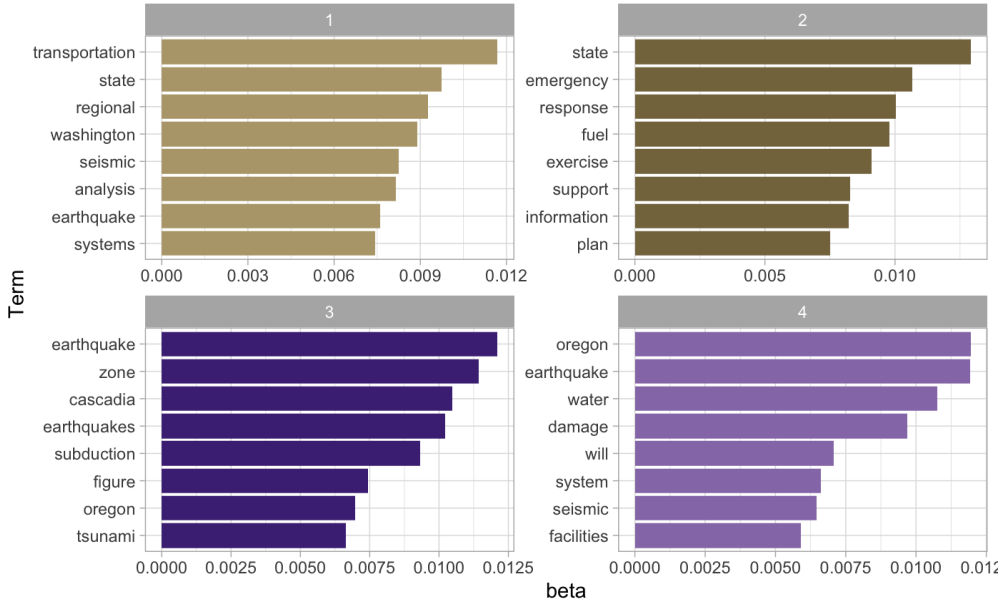


Figure 2.8: Topic modeling with four topics using LDA. The user specified a number of topics and documents that are categorized by shared terms and themes.

on the Portland metropolitan region's economics, emphasizing transportation, and helping readers identify relevant documents. This type of visualization is generally helpful in summarizing the main themes of the corpus at a high level and in identifying similarly themed documents. For example, the documents in Topic 2 are generally state-level response exercise action-action reports or plans, whose cross-referencing is helpful in practice, e.g., state-level benchmarking.

The potential for applying topic modeling reaches beyond disaster preparedness reports, benefiting any domain where analyzing large document sets for key themes is essential. By clustering related documents and highlighting the most pertinent information, this tool streamlines knowledge management and allows professionals to grasp central themes efficiently.

Topic-word scores using Bidirectional Encoder Representations from Transformers (BERT) - what are the most common topics discussed in this corpus?

For this analysis, no preprocessing of documents is done, and stop words remain intact.

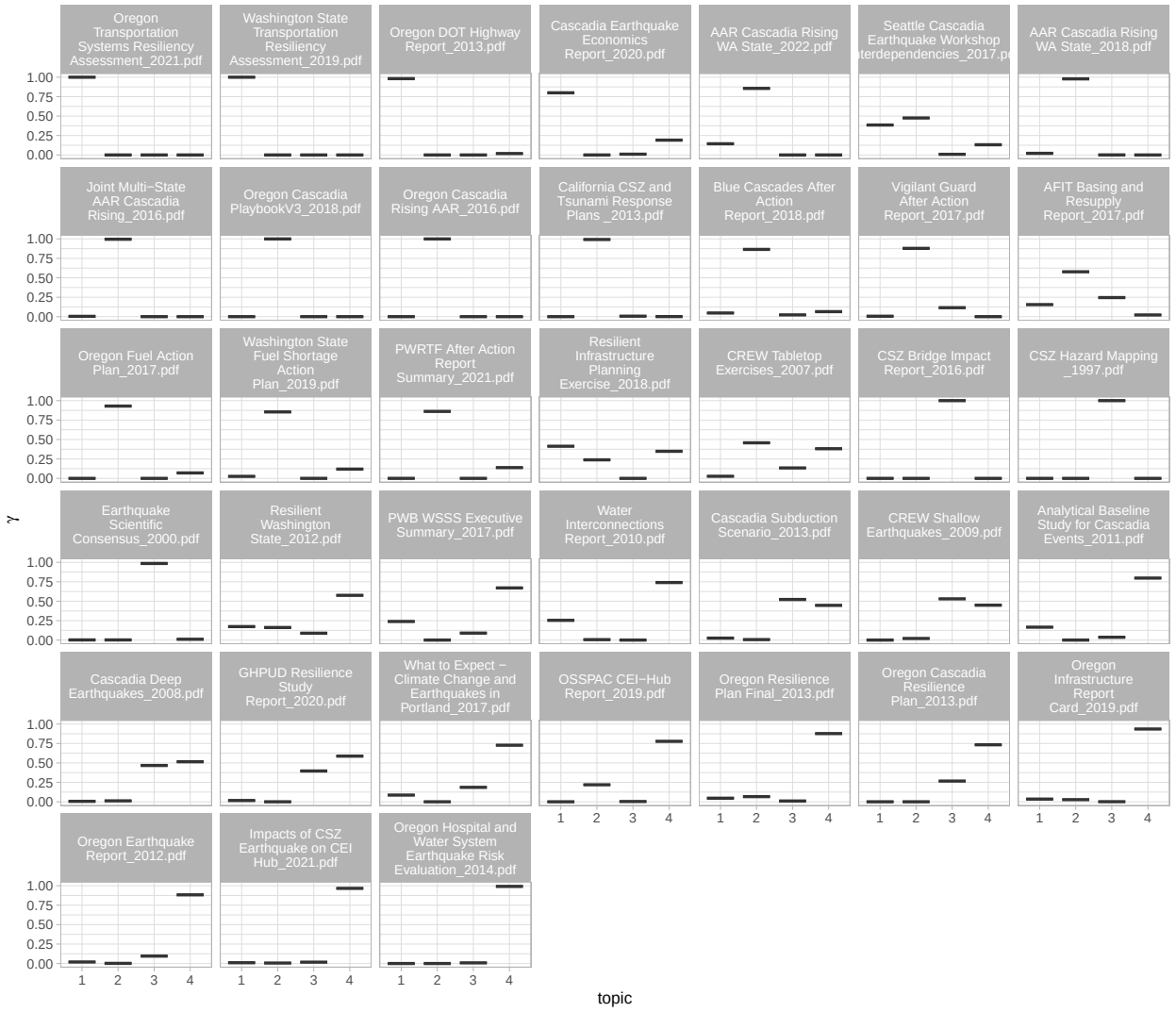


Figure 2.9: Document-topic probabilities using LDA. Quantitative representation of how each document in the corpus relates to the identified topic. (See Figure 4 in Appendix A for an alternative visualization of which documents contribute to which topics.)

Figure 2.10 represents the class-based tf-idf of each representative topic identified by `BERTopic`, which leverages BERT and c-tf-idf reviewed in Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT). `BERTopic` allows for a flexible choice of clustering algorithms for group-related terms. We chose Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) because of its ability to select an optimal number of topics (i.e., clusters) by maximizing a measure of the overall stability of clusters [16]. The resulting five topics of `BERTopic` show good resemblance with the four topics chosen by the LDA in Topic modeling with four topics using Latent Dirichlet Allocation (LDA) - what are the most common topics discussed in the corpus?: Compare the `BERTopic`'s Topics 0, 1, 2, 3, and 4 with the LDA 4-Topic model's Topics 1, 2, 3, 4, and 2, respectively. In other words, `BERTopic` found similar topics to LDA, except for further breaking down LDA's Topic 2 into Topics 1 and 4. In sum, LDA and `BERTopic` identified consistent topics within the corpus, demonstrating their robustness. However, other corpora may lead to substantially different topics being determined by these two fundamentally different tools.

Short topic description using GPT-3.5 - how can you summarize the content of representative documents in each identified topic in Figure 2.10?

GPT-3.5 is a version of OpenAI's Generative Pre-trained Transformer model, which is specifically designed for natural language processing[81]. For each topic identified by the `BERTopic` model earlier, the short description in Table 2.2 and the one-paragraph summary in Table 2.3 provide a fuller picture of each topic's representative documents in conjunction with each topic's keywords in Figure 2.10. A summary of six outlier documents is provided as well. To create the short description in Table 2.2, we prompted the GPT model with "I have a topic that contains the following documents [DOCUMENTS]. The topic is described by the following keywords: [KEYWORDS]. Based on the above information, can you give the topic a short label?" To create the paragraph summary in Table 2.3, we prompted the GPT model with the same [DOCUMENTS] and [KEYWORDS] prompt as with the short description but added a further prompt of "Based on the information above, please give a 100-word description of this topic".

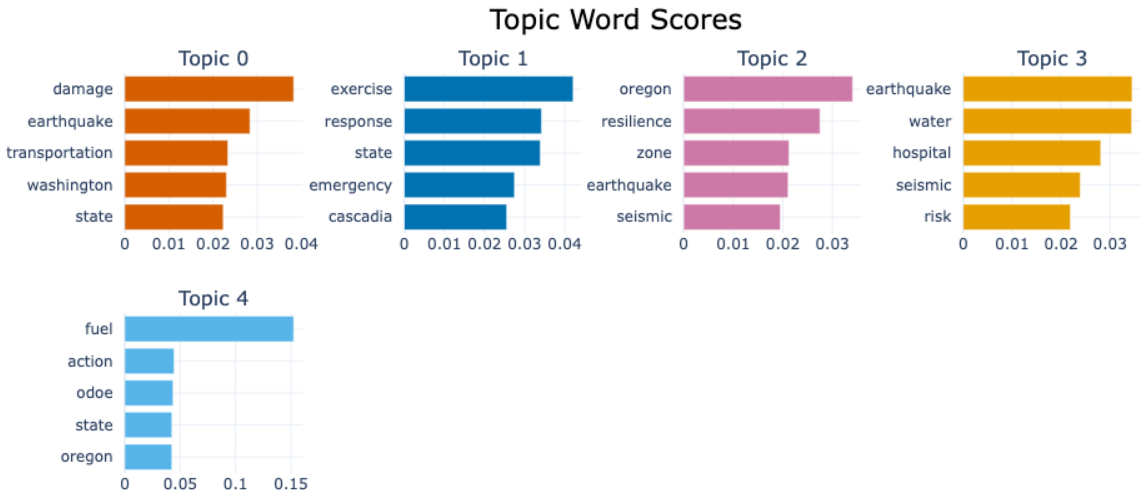


Figure 2.10: Topic-word scores using Bidirectional Encoder Representations from Transformers. Term frequency - inverse document frequency (c-TF-IDF) term scores by topic. The score reflects important terms within the identified topic cluster.

The generated summaries or representations are overall helpful in gaining a quick overview of the topics. However, the artificial intelligence (AI)-generated wordings (i.e., AI's best attempt to predict a sequence of words) in response to the prompt can be misleading. For example, Table 2.2's Topic 0 representation can mislead uninformed readers to believe that all eight representative documents are part of the Resilient Washington State Initiative. Similarly confusing is Topic 2's word choice "Oregon Resilience Plan," which is specific to one document from 2013, representing all seven documents in the topic. The longer-form representations in Table 2.3 are more informative overall but plagued by factual errors and misrepresentations (e.g., Topic 2's "Oregon Resilience Zone" is the AI's made-up term, and Topic 4's references to the documents' last revised timestamps are not representative of all four documents). This issue of problematic AI products is an active research topic and at the center of ongoing societal debates [102]. Further clarification of these summaries may also be seen with the refinement of provided keywords, through the use of other keyword and sentence extraction algorithms such as TextRank [67] or Rapid Automatic Keyword Extraction [91].

Topic	Count	Representation
-1	6	Oregon's Infrastructure and the Impact of Cascadia Subduction Zone Earthquake
0	8	Resilient Washington State's Infrastructure Assessment and Recovery Framework
1	7	Cascadia Rising Exercise Response and Planning
2	7	Oregon Resilience Plan for Cascadia Subduction Zone Earthquake and Tsunami
3	6	Earthquake and Seismic Risk Assessment for Oregon's Critical Infrastructure
4	4	Fuel Action Plan in Oregon for Emergency Water Supply

Table 2.2: Sentence topic description using GPT-3.5. A topic summary was generated using GPT-3.5 on the keywords and representative documents from the `BERTopic` model. Topic (-1) includes outliers. The count is the number of representative documents. Representation is the summarization of the representative documents.

2.5.5 Synthesis of case study findings: practical implications for emergency managers

In the section above, we discussed findings specific to each tool and its associated probing question. However, several overarching themes emerged from this analysis.

Topic	Count	Representation
-1	6	This topic focuses on the potential impact of a Cascadia Subduction Zone earthquake on various aspects of Oregon. The documents highlight the evaluation of infrastructures and the necessary upgrades to withstand such earthquakes. Additionally, there is information about Cascadia Subduction Zone earthquakes, including a scenario of a magnitude 9.0 earthquake. The topic also includes a summary of a Penrose Conference that discussed the tricentennial anniversary of the Great Cascadia Earthquake. Overall, the topic revolves around understanding the hazards posed by earthquakes in Oregon's Cascadia Subduction Zone and the measures that can be taken to mitigate their effects on infrastructure and public safety.
0	8	This topic focuses on the assessment and analysis of the impacts of earthquakes, tsunamis, and other natural hazards on transportation infrastructure in the Pacific Northwest region, specifically in Washington State and Oregon. The documents mentioned highlight the importance of assessing the resiliency and vulnerability of regional bridge systems and port facilities in the face of potential seismic events, such as the Cascadia Subduction Zone earthquake. The analysis includes considerations of damage, emergency response scenarios, liquefaction risk, and the expected impacts on transportation systems and facilities. The goal is to develop strategies and plans to enhance the resiliency and preparedness of the infrastructure in the region.
1	7	This topic focuses on the planning and response efforts for a potential catastrophic earthquake and tsunami event in the Cascadia Subduction Zone. The documents mentioned include reports on the Cascadia Rising 2016 exercise, which involved multiple states and jurisdictions, including California and Washington. These reports discuss the coordination and management of information, operations, and resources at local, state, and federal levels. The goal is to enhance the capability and readiness of agencies and organizations involved in emergency response and recovery. The importance of public awareness, support, and collaboration in mitigating the impact of such a disaster is also highlighted.
2	7	This topic focuses on the Oregon Resilience Zone and its preparation for a potential earthquake in the Cascadia subduction zone. It discusses plans and strategies to enhance the resilience of buildings, infrastructure, and critical facilities, particularly in coastal areas. The topic also highlights the importance of mitigating fuel and oil-related risks and the need for recovery and recovery models. Documents such as the Oregon Resilience Plan and the CEI Hub Mitigation Strategies provide insights into the state's efforts to reduce risk and improve recovery in the event of a catastrophic earthquake and tsunami. The topic emphasizes the criticality of preparedness and response measures in mitigating the impact of such a natural disaster.
3	6	This topic focuses on the seismic risk and impact of earthquakes in the state of Oregon, particularly in relation to critical infrastructure such as hospitals and water systems. The documents highlight the economic analysis of a potential Cascadia subduction zone earthquake in the Portland metropolitan region and the risks faced by Oregon's critical energy infrastructure hub. It also includes a study on the earthquake risk evaluation of hospitals and water systems in the state. The topic covers various aspects such as assessing damage and risks, energy transmission systems, and the potential impact on the city of McMinnville and the Lincoln and Portland areas.
4	4	This topic focuses on fuel action and emergency response plans related to the state of Oregon. It includes keywords such as fuel, action, emergency, water, supply, agencies, and support. The mentioned documents highlight the importance of preparedness and response in the face of fuel shortages and the need for collaboration among water providers in the Portland metropolitan area. The topic also involves assessing information, planning, and coordinating efforts at the county, federal, and agency levels, with a primary mission to ensure an adequate fuel supply, especially during emergencies. The documents referenced were last revised in October 2017 and April 2019.

Table 2.3: Paragraph topic description using GPT-3.5. A topic summary was generated using GPT-3.5 on the keywords and representative documents from the `BERTopic` model.

Communicating risk of M9 events

Our corpus of practical reports comprises publicly available reports aimed at raising community and professional awareness of the potential impacts of a Cascadia subduction zone M9 megathrust earthquake. Our analysis has identified several opportunities for enhancing effective communication with the target audience. This includes uncovering latent sentiment expressed in the corpus (its predominant negativity shown in Figure 2.6) and observing inconsistencies in vocabulary usage as shown in Term frequency - inverse document frequency of whole corpus - what are each corpus document's unique/characteristic themes?. These findings underscore the importance for authors–whether practitioners or not–to be intentional in preparing practical reports, ensuring that they convey their message effectively. This requires considering the intended audience, the content being communicated, and how it is presented. Effective communication of disaster preparedness information can impact how communities, agencies, and citizens receive, process, and react to information [52].

Characterizing priorities

Many of the tools described in this work aided in understanding the priorities of the represented entities. Even during the initial phase of corpus development, our Google search revealed potential disparities among states. Oregon exhibited a more public-oriented approach to readiness, making its literature readily accessible and tailored towards individual preparedness. Conversely, Washington and California had fewer publicly available resources and appeared to focus more on documenting the outcomes of readiness exercises. Furthermore, our analysis extracted the potential priorities of practitioners, as evidenced by the prioritization of critical infrastructures depicted in Figure 2.2 and discerned differences in priorities between Washington and Oregon state-level resilience plans as shown in Figure 2.5. We also identified shifts in priorities over time, exemplified by the comparison of Cascadia Rising AARs in 2018 vs. 2022 in Figure 2.5. Understanding the priorities of stakeholders within the emergency management field is paramount, as it can hold critical implications for the efficacy and efficiency of essential response tasks (e.g., managing tradeoffs and limiting conditions [83]). In harnessing the power of text mining, organizations can enhance communication and

coordination efforts to ensure streamlined response efforts.

2.5.6 Brief survey of emergency management researchers and practitioners

We asked over 100 emergency management researchers and practitioners for their voluntary, anonymous participation in reviewing our 9-page white paper (found in Appendix A) and completing our 2-question online survey (Question 1 - "What specific aspects of the tools did you find particularly useful? Why?" Question 2 - "What challenges or limitations do you see in the tools shown in this document? Why?") Of those surveyed, we received five responses from researchers and 22 responses from practitioners. Their feedback is summarized in Table 2.4. Though participants were not explicitly asked to describe their previous knowledge of the tools presented, many indicated they had little to no prior experience with them.

Researchers ($n = 5$)	
Specific tools found to be most beneficial	• Term Frequency-Inverse Document Frequency of Select Pairs [Term frequency - inverse document frequency of select pairs - how do characteristic terms in one document compare to those of another related document?] • Sentiment Analysis [Sentiment analysis - what is the portrayed sentiment of this corpus?] • Topic Modeling w/4 Topics Using LDA [Topic modeling with four topics using Latent Dirichlet Allocation (LDA) - what are the most common topics discussed in the corpus?] • Document Topic Probabilities Using LDA [Document-topic probabilities using Latent Dirichlet Allocation (LDA) - which documents comprise the identified topics in Figure 2.8?]
Potential applications/benefits	• Ability to understand a global perspective for policy implementation • Ability to understand prevalent framings and gaps in the documents that comprise the corpus • Ability to objectively visualize documents to support equitable decision-making • Ability to identify specific document(s) related to a given topic to facilitate efficient review

Limitations and challenges	• Ease of use and training for implementation • Accuracy of AI-generated summaries • Need for expert input/review of results to ensure precision • Corpus curation/development
Practitioners ($n = 22$)	
Specific tools found to be most beneficial	• Term Relationships [Term relationships - what are the predominant themes in this corpus?] • Term Frequency-Inverse Document Frequency of Select Pairs [Term frequency - inverse document frequency of select pairs - how do characteristic terms in one document compare to those of another related document?] • Sentiment Analysis [Sentiment analysis - what is the portrayed sentiment of this corpus?] • Sentiment Frequency [Sentiment frequency - what emotions are portrayed most in a single document (Oregon Resilience Plan 2013)?] • Topic Modeling w/4 Topics Using LDA [Topic modeling with four topics using Latent Dirichlet Allocation (LDA) - what are the most common topics discussed in the corpus?] • Short Topic Description Using GPT-3.5 [Short topic description using GPT-3.5 - how can you summarize the content of representative documents in each identified topic in Figure 2.10?]
Potential applications/benefits	• Ability to account for and represent human emotion in corpus • Ability to identify priorities, strengths, and weaknesses from AARs for action plans • Ability to identify gaps in documents that comprise the corpus and gaps in coverage of relevant topics for the intended audience • Ability to present comprehensible visualization • Use of tool to determine where best to focus efforts (e.g., deep dive into certain document(s))
Limitations and challenges	• Ease of use and training for implementation • Accuracy of AI-generated summaries • Corpus curation/development

Table 2.4: Summary of survey results from emergency management researchers and practitioners.

Overall, researchers tended towards more traditional and proven text mining approaches, such as LDA and tf-idf, instead of the more state-of-the-art approaches, noting their hesitancy for trusting AI-based tools. They saw value in applying text-mining tools to visualize large amounts of data to

support policy implementation and decision-making processes.

Practitioners found sentiment analysis and sentiment frequency to be the most intriguing and potentially valuable tools for their work. More than half of the respondents noted that understanding the human emotions surrounding a document or corpus could contribute to their work and bring added value to their understanding of AARs, action plans, and assessments. Many respondents also noted that using feature analysis (tf-idf and term relationships) to compare various documents would be particularly helpful (as shown in Figure 2.5), as it allows a researcher to see a shift or change in priorities over time quickly.

Both researchers and practitioners expressed concerns about the cost-benefit tradeoffs in time and expertise needed to learn and implement new tools. They also noted that results heavily rely on the quality of input, placing heavy emphasis on developing and curating the studied corpus.

2.6 Discussion

In this section, we discuss the strengths and limitations of our proposed approach, identify practical implications for researchers and practitioners, and discuss future work.

2.6.1 Approach strengths and limitations

Practical reports as a corpus

This paper introduces a corpus of practical reports on the Cascadia megathrust earthquake, analyzed using text mining tools. These reports offer authentic insights into planning and preparedness for this event, grounded in real-world experiences. Researchers and practitioners can uncover valuable patterns, trends, and relationships by applying text mining tools to this corpus, enhancing understanding and informing decision-making. Text mining enables a comprehensive analysis beyond individual report examination, providing generalizable insights into the challenges and opportunities identified across the corpus.

One noted limitation is that the quality of text mining output depends on the quality of input data. We systematically collected practical reports for analysis, but were restricted to publicly available documents. Furthermore, the structured nature of practical reports may limit the depth and

richness of information compared to more unstructured sources. Despite these limitations, using tools that ensure equitable comparison, as demonstrated in Figure 2.5, and integrating domain-specific knowledge into the analysis can mitigate these challenges. Viewed through a different lens, pooling similar (though not identical) reports for analysis maintains confidentiality and proprietary information while extracting insightful thematic information from the corpus. Moreover, the structured format of these reports could be considered advantageous for future work, facilitating targeted mining for specific objectives, methods, results, and more.

Text mining as an approach

In this work, we introduce text mining as a powerful tool for efficiently extracting insights and knowledge from large volumes of textual data. It is flexible and scalable so that users can go beyond the surface level of the text and discover insights such as sentiment/emotion and latent topics, and can be applied to a variety of corpora. Text mining can empower researchers and practitioners to harness the information contained within textual data, driving innovation, decision-making, and discovery across various fields and applications.

A recognized drawback of using text mining tools is the requirement for domain-specific and background understanding of the corpus to enhance interpretability and identify potential misrepresentations. Human language is inherently subjective and complex, making it essential for users to possess familiarity with the subject to navigate nuances effectively [22]. For example, one needs to be informed on the disaster preparedness field and its language to identify topic themes in topic modeling (e.g., recognizing an overarching theme encompassing the characteristic terms identified for each topic from a corpus) or decipher varying trends among compared reports (e.g., comparing WA vs. OR resilience plans. As shown in AI-assisted topic modeling, domain knowledge is also required to validate results (e.g., identification of "Oregon Resilience Zone" as a made-up term). Therefore, despite the promise of text mining and AI, in general, to help users learn from a corpus, those with experience in the field of the corpus are better positioned to benefit from the tools while guarding against their risks (e.g., misinformation from AI).

Furthermore, text mining, which offers the convenience of analyzing a large amount of text

data computationally, is meant to *complement* more in-depth, targeted research tools, which tend to require more resources (e.g., subject-matter experts' time to conduct qualitative analysis). For example, text mining-based exploratory analysis may help form specific research hypotheses and questions that may be probed through follow-up qualitative research.

Suite of tools

Many text mining tools are available, but in this work, we opted for ones that offered the most value for their usability and accessibility. The tools presented are easy to access through open-source programming languages (`R` and `Python`) utilizing user-friendly packages like `tidytext`, `ggplot2`, and `BERTopic` to conduct analysis and produce high-quality, concise visualizations. The implementation of these tools requires basic computer programming skills. Still, with a modest investment of time, users can generate meaningful products tailored for use with any identified corpus.

Each text mining tool has its own set of strengths and limitations. Potential users may find Table 2.4 helpful in identifying the appropriate tools.

2.6.2 Practical implications for researchers - bridging the knowledge gap

The research-practice knowledge gap is not a novel challenge, but an enduring dilemma that transcends various disciplines, persisting despite ongoing efforts to bridge the divide. In this study, we present a novel approach - text mining of practical reports - to objectively demonstrate a practitioner's perspective as embodied in these reports. While researchers are adept at understanding the research perspective based on their experience and expertise, they may lack some insight into practitioners' viewpoints. By leveraging our research-based tools, even those with limited coding experience can utilize their seasoned expertise to identify discrepancies between research and practitioner perspectives.

2.6.3 Practical implications for practitioners - a valuable addition to your toolkit

The integration of text mining presents practitioners with a valuable augmentation to traditional research tools and enables them to gain objective insights from diverse corpora efficiently. By em-

Document and Corpus Feature Analysis		
Term Frequencies	(+)	simple and straightforward approach of counting words' frequencies to identify predominant themes in a corpus or document
	(–)	lack of ability to determine reasoning/meaning behind included and/or missing terms
	(–)	lack of calibration for the text of varying lengths (e.g., longer texts = greater frequency of words)
Term Relationships	(+)	simple and straightforward tool to count the frequency of sequential and adjacent terms to identify patterns, themes, and relationships between terms
	(+)	can be used for corpus or single-document analysis
	(–)	lack of ability to determine reasoning/meaning behind included and/or missing terms and relationships
	(–)	lack of calibration for texts of varying lengths
Term Frequency-Inverse Document Frequency (TF-IDF)	(+)	simple and straightforward tool to analytically identify the important words in each document by placing heavier weight on the words that are used less frequently in each document
	(+)	allows for comparison of documents of varying length
	(+)	easily meldable for use in other tools
	(–)	can create unwieldy large vector spaces for documents with large vocabularies
	(–)	does not account for similar words within vocabularies
	(–)	lack of ability to determine semantic usage of terms
Sentiment Analysis		
Lexicon-based Sentiment Analysis	(+)	easily accessible use of pre-established dictionaries to extract the emotional polarity of text
	(+)	can be used for corpus or single-document analysis
	(–)	general-use lexicons may represent different sentiments than intended by the author – may require domain-specific lexicon
Topic Modeling and Summarization		
Latent Dirichlet Allocation (LDA)	(+)	highly interpretable probabilistic model assigning documents to topics
	(–)	number of topics must be specified – hyperparameters require tuning
`BERTopic` and GPT Models	(+)	pretrained word embedding algorithm that allows for semantic relationships
	(+)	results are easily interpreted
	(+)	number of topics is determined automatically by an algorithm
	(–)	word embeddings can make this model very large, requiring dimension reduction
	(–)	carries general risks of AI (e.g., misrepresentation)

Table 2.4: Strengths and limitations of text mining tools presented in this work. (+) denotes a strength, and (–) denotes a limitation.

ploying this suite of tools, practitioners can access impartial insight into a corpus, which they may choose to complement conventional approaches such as literature review, environmental scans, landscape analysis, or a structured checklist. For example, emergency management consultants have the opportunity to customize the suite of tools provided, enabling them to seamlessly incorporate new corpora for analysis across different clients and alongside their conventional manual analyses. Text mining offers an objective lens, which the reader's presuppositions might subjectively influence. Emergency managers and other busy practitioners who lack time to review numerous reports individually can benefit from the suite of text mining tools. While some initial preparatory work is necessary, once implemented, the suite facilitates a much quicker, high-level overview of the corpus than traditional approaches (e.g., reading through the corpus). Incorporating text mining tools empowers practitioners to elevate their research approaches and attain unbiased insights across various corpora. These insights can then be used as hypotheses to inform and spur further research, and then, once tested, encourage potential actions in practice.

2.6.4 Future work

This work offers many expansion opportunities. Researchers can build on the presented toolkit by broadly exploring emerging tools in text mining, natural language processing, and AI. Further application of these tools to other corpora of practical reports can deepen our understanding of the advantages and limitations of the tools and expand their applicability and accessibility to researchers and practitioners. Our proposed approach allows them to extract overarching themes and novel insights from a corpus. In doing so, additional research questions are bound to emerge (e.g., the impact of sentiment on mitigation and preparedness behaviors). Insights gained from the presented tools can spark other researchers and practitioners to formulate follow-on questions, objectives, and potential actions to advance their respective fields.

Chapter 3

PREDICTING KEY BRIDGE CHARACTERISTICS USING MACHINE LEARNING

3.1 Abstract

We explore the use of machine learning (ML) to predict key structural characteristics of bridges using data from the U.S. National Bridge Inventory (NBI). The publicly available NBI (2023) documents more than 100 characteristics of approximately 617k bridges in the United States, but few directly relate to structural or earthquake engineering. To provide ground-truth values, key structural characteristics not documented in the NBI were extracted from structural drawings of 796 bridges within Washington State. Using the corresponding NBI characteristics (features) and extracted properties (targets), we trained and evaluated standard ML algorithms and an automated ML (AutoML) approach using AutoGluon to develop predictive models for four target characteristics. We demonstrate that ML can feasibly generate a more detailed and comprehensive bridge inventory, supporting regional-scale planning, resource prioritization, and post-earthquake response. More specifically, we found that while prediction accuracy varied by target, ensemble methods consistently delivered the best results. To demonstrate the application of this work, we applied the best-performing algorithm to predict the capacity-to-demand ratio for large-deformation, cyclic, shear failure of bridge columns, a critical characteristic of bridge seismic vulnerability. As expected, the year of construction was the feature that contributed the most to the model, because it correlates well with the amount of transverse reinforcement. Still, the incorporation of additional features further improved its predictions. The predicted characteristics were accurate enough to improve the identification of shear-critical columns. The practical value of quick identification lies in enhancing resilience by enabling rapid decision-making, such as prioritizing pre-earthquake mitigation and post-earthquake inspections, particularly when reviewing plans for every bridge in a region is not feasible.

3.2 Introduction

Bridges serve as critical links in the transportation system after disasters caused by natural hazards, providing essential connections for transportation, emergency response, and supply chains. Their functionality directly impacts the speed and effectiveness of short-term response and long-term recovery efforts. However, the lack of comprehensive knowledge of bridge characteristics, performance, and network dynamics during or after extreme events severely limits realistic simulation and modeling of bridge systems at the regional level.

Currently, the only national database to monitor the condition and safety of bridges is the National Bridge Inventory (NBI). It contains over 100 elements related to location, structural attributes, condition ratings, traffic data, and historical information about the approximately 617k bridges in the United States [2]. This information is essential for allocating federal funds to support prioritized maintenance, rehabilitation, and replacement of bridges based on routine usage, scheduled inspection records, and generalized maintenance plans.

Despite this, the NBI lacks detailed information on bridges' seismic vulnerability, even though these structures are critical components of transportation networks in earthquake-prone regions. Collecting this additional information often requires on-site bridge inspection, thorough review of bridge plans, and maintenance of large databases. Such a detailed characterization of the bridge inventory would be extremely expensive and time-consuming. For example, California, Oregon, and Washington have over 32,000 bridges. Conducting comprehensive seismic assessments for these structures would demand significant financial and personnel resources, underscoring the urgent need for scalable, data-driven methods to supplement or prioritize traditional approaches.

However, the Washington State Department of Transportation (WSDOT) and the University of Washington (UW) collected detailed bridge information in the WSDOT inventory over several years. This Key Bridge Characteristics (KBC) dataset includes 120 characteristics related to seismic vulnerability of approximately 800 bridges, including information on the substructure, including piers, abutments, bearings, shear keys, and foundations, which are often critically important when predicting long-term bridge condition and bridge performance in natural hazards.

Machine learning (ML) may enable using individual bridge information in the NBI and the small

set of training data available in the KBC to predict other bridge characteristics better aligned with bridge condition ratings and bridge performance in an earthquake. These predicted inventory characteristics may then be used in existing models that predict, for example, long-term bridge condition, bridge seismic performance, and bridge performance in flooding and storm surge. This information is vital for DOT planning tasks, including budget planning for maintenance and replacement, retrofit for functionality, and emergency operations planning. Additionally, predicting these other bridge characteristics at a larger scale enables more integrated and cross-coordinated planning efforts (e.g., regional transportation system modeling, regional emergency route planning, regional prioritization for repair/replacement efforts while minimizing network disruption).

Our main contribution is demonstrating the feasibility of using ML to generate a more detailed and comprehensive database of bridge characteristics that can then be used to enhance current analyses on a regional scale. We do this by using a variety of ML algorithms to predict a set of bridge characteristics related to seismic vulnerability contained in the Key Bridge Characteristics dataset using National Bridge Inventory data. This approach reduces the time and labor resources required to collect this information manually. It increases the inventory of bridge characteristics available for use in regional-level transportation modeling and analysis efforts.

As a secondary contribution, we present a practical application of our extended inventory by leveraging the best predictions of key bridge characteristics to calculate the *Shear-Critical Column Index (SCCI)* to identify bridges susceptible to shear failure. The insights derived from this analysis can assist engineers in prioritizing post-earthquake inspections of shear-critical columns and strategically allocating limited retrofit resources to bridges at the highest risk of column shear failure.

3.3 Prior machine learning applications using NBI data

Machine learning (ML) has evolved rapidly over recent years and continues to grow in application areas, including engineering. ML is well-suited for engineering problems because it can efficiently handle complex problems, but it usually requires large amounts of high-quality data to increase effectiveness [108]. The NBI provides comprehensive data on nearly every publicly-owned bridge

in the United States, follows federal guidelines for collection, and is regularly updated. This large, high-quality dataset has been used in many ML capacities to predict future bridge condition ratings and inform maintenance prioritization. For instance, Mia and Kameshwar [66] employed ML techniques from historical NBI data to forecast the future deterioration of key bridge features. In contrast, Alogdianakis et al. [6] and Alipour et al. [4] explored deterioration patterns in the NBI using ML. Fiorillo and Nassif [37] and Liu and Zhang [62] leveraged NBI data to predict future condition ratings of bridge elements, and Gungor et al. [44] combined weight-in-motion (WIM) and NBI data with ML to predict deck conditions and determine overweight vehicle fees. ML and the NBI have also been used as a preliminary design aid for engineers in selecting the optimal bridge type [53]. The significance of the NBI is underscored by efforts to clean and enhance the data, as demonstrated in the work by Fereshtehnejad et al. [36] and Hu and Assaad [48]. The variety of machine learning applications in engineering highlights the suitability of using the NBI for our machine learning work.

3.4 Algorithms

When predicting key bridge characteristics, we began our analysis by exploring a set of time-tested algorithms that perform generally well across many datasets. This manual exploration of the data allowed us to analyze the data and understand the problem better while supporting the execution of domain-specific data preprocessing tasks that could be overlooked in more automated methods. We implemented these singular algorithms using regressors from `Scikit-learn` [85], a widely used `Python` library that provides efficient and well-documented tools for predictive modeling. This ensured consistency in model implementation and allowed for effective benchmarking before integrating a more automated approach.

This section briefly overviews the general category of algorithms used in our analysis.

Decision trees. Decision trees are easy to implement and interpret and can handle a mix of numerical and categorical data. However, they tend to overfit and are sensitive to minor variations in the data [14]. Given the novelty of this work, this algorithm was well-suited for some initial,

exploratory work on our dataset.

Random forests. Building on our decision tree analysis, we explored random forests because they tend to achieve a higher accuracy rate and can more effectively handle missing data, thereby reducing overfitting[93]. Despite limited interpretability, we chose this algorithm for its robust performance across various datasets.

Neural networks. Though neural networks typically perform better with larger datasets, we chose to implement this algorithm for its ability to find complex patterns and relationships in data [28]. This algorithm can also be challenging to interpret due to its 'black box' nature [115], but its advantages in capturing nonlinear dependencies outweigh the limitations.

Gradient boosting. We employed gradient boosting, specifically `XGBoost`, for its robust performance and efficiency in handling structured data to optimize model accuracy. Also, its capability to handle missing values and mitigate overfitting made it a reliable choice for regression and classification tasks [21].

Support vector machines. We selected support vector machines for their effectiveness in classifying linear and nonlinear tasks, particularly their ability to create optimal hyperplanes that maximize the margin between different classes. Their strong performance in high-dimensional spaces makes support vector machines suitable for our dataset, despite potential challenges associated with kernel parameter selection and class separability [77].

Automated machine learning (AutoML). AutoML represents a significant advancement in the field, automating various ML tasks to make the technology more accessible to non-experts while enhancing efficiency for seasoned practitioners [49]. By leveraging Auto ML algorithms, such as AutoGluon [30], the entire ML workflow was streamlined, encompassing crucial processes such as model selection, hyperparameter tuning, feature engineering, and model ensembling.

To predict key bridge characteristics, we implemented AutoGluon and compared performance

with standard algorithms. In characterizing shear-critical columns, we utilized AutoGluon in our modeling work to predict the *Shear-Critical Column Index*.

3.5 Primary data sources

3.5.1 Key Bridge Characteristics (KBC) dataset

Bridge engineers and students from the Washington State Department of Transportation and the University of Washington collected 120 characteristics for 796 highway bridges located in the Puget Sound Region and on the Olympic Peninsula in Washington State. The resulting dataset documents key characteristics of the bridge locations, superstructures, shortest piers, tallest piers, and end abutments related to seismic vulnerability. This information was collected by manually examining bridge plans contained in the WSDOT Bridge Engineering Information Systems (BEISt) and documenting measurements in the database.

This research targets bridges most susceptible to shear failure, focusing specifically on multi-span structures supported by reinforced concrete columns with constant cross-sections. Of the 796 complete bridge records analyzed, 37 were supported by walls, 112 by piles, 45 by tapered columns, and 115 were single-span bridges. The remaining 478 bridges, approximately 60% of the dataset, met the criteria for detailed analysis, consisting of multi-span structures with non-tapered reinforced concrete columns. The distribution of support types is shown in Figure 3.1. The inventory is predominantly composed of spiral-reinforced columns with circular cross-sections (76%), followed by rectangular columns with rectangular hoops (10.5%), rectangular columns with spiral reinforcement (7%), oblong cross-sectional columns (5%), and a small number of columns with hexagonal cross-sections (1.5%).

Figure 3.2 shows the distribution of the construction year for these bridges. Most (62%) of the bridges were built between 1960 and 1974, corresponding to the peak construction years of the interstate highway system in the United States [32]. Only 28% of bridges were built after 1975, when bridge codes began to be modified to reflect lessons learned during the 1971 San Fernando earthquake. For example, Figure 3.3 shows the relationship between the spacing of the transverse reinforcement and the construction year for the considered bridges. Before the mid-1970s, the

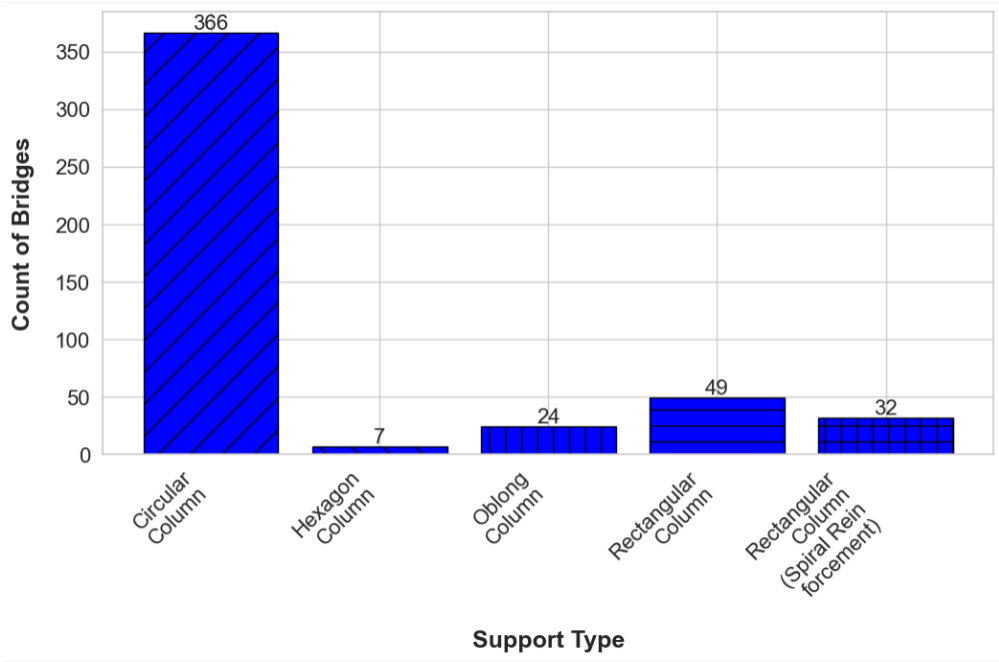


Figure 3.1: Count of bridges by support type

spacing of the transverse reinforcement was often specified as 12 inches (305mm), a value based on the need to hold longitudinal bars in place, rather than on the needs for concrete confinement or column shear strength[10].

3.5.2 National Bridge Inventory (NBI)

The NBI is a comprehensive database maintained by the Federal Highway Administration (FHWA). It includes details such as bridge identification, structural information, condition ratings, load capacities, and traffic data for over 617,000 publicly owned highway bridges in the U.S. State and federal agencies update it annually, following submission guidelines, and it is made available each June, with periodic updates throughout the year. For this study, we used the 2023 NBI data[33].

The NBI contains 133 features for 8,474 bridges in Washington State. Despite submission standards, some data fields are incomplete. We removed columns with significant missing data, reducing the dataset from 133 to 95 features without losing essential information. We also removed irrele-

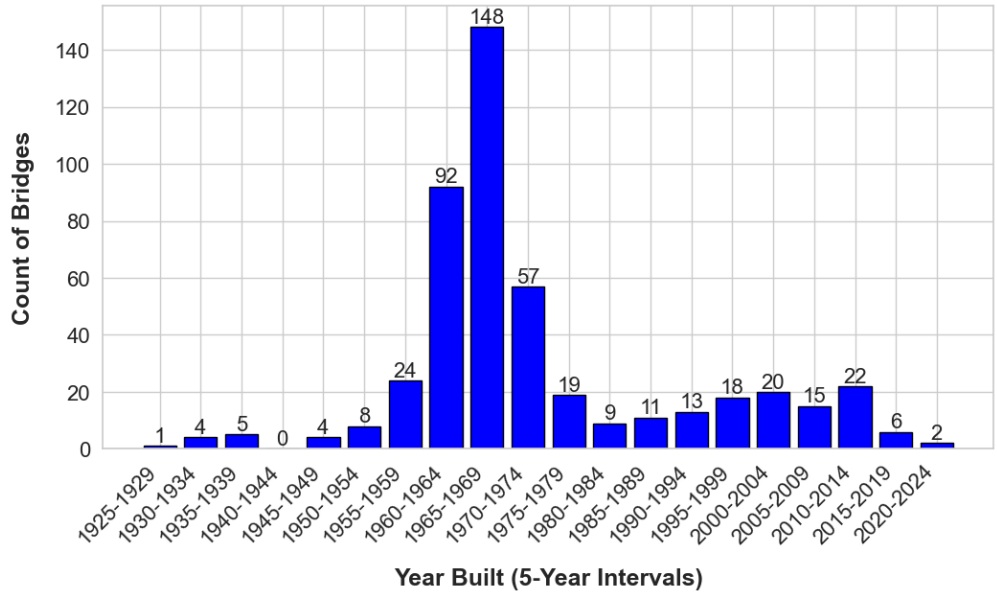


Figure 3.2: Count of bridges by year of construction

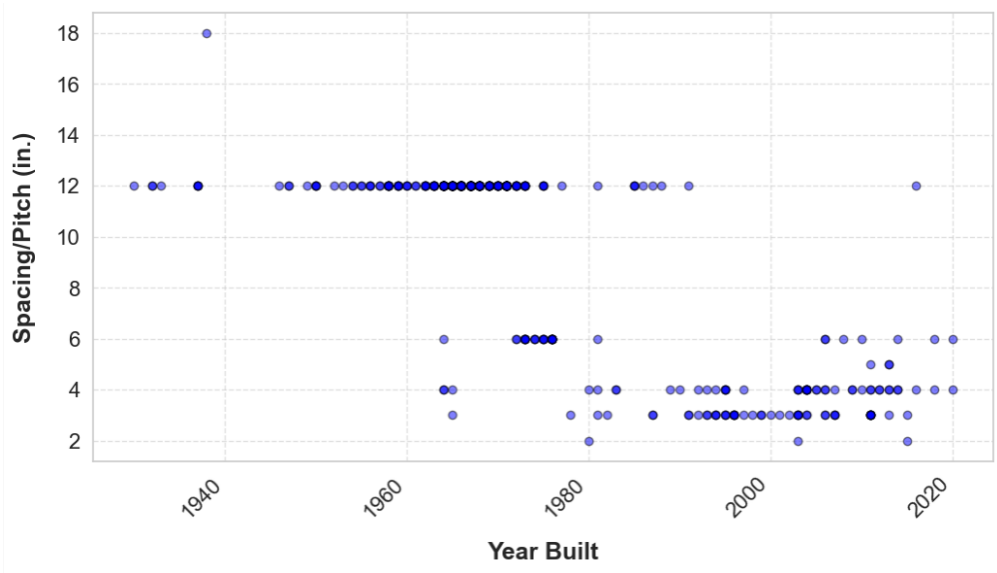


Figure 3.3: Relationship between center-to-center spacing between spiral or ties on bridge column and year built

vant features, such as those with unique values for all entries (e.g., bridge identification numbers) or constant values (e.g., state code). The retained 95 features comprise numeric measurements, nominal and ordinal categorical variables, and binary indicators.

3.6 Predicting key bridge characteristics using machine learning

This section details the modeling approach illustrated in Figure 3.4. We begin by outlining the database preparation process, followed by a discussion of model training, evaluation, and selection. Finally, we present and interpret the results.

3.6.1 Data preparation

Of the 120 characteristics in the Key Bridge Characteristics dataset, we identified an initial set of 4 target variables for prediction. These variables, listed in Table 3.1, are measurements from the bridge's substructure and are directly related to the seismic shear vulnerability of the reinforced concrete columns. A histogram for each target variable's distribution is shown in Figure 3.5.

Target Variable	**Description**	**Median**	**Training Size**
Columns per Bent [1-8]	Number of columns supporting the shortest bent	3.00	360
Column Minimum Aspect Ratio (L/H_{MIN}) [0.46 - 13.08]	Ratio of column height to its least lateral dimension	5.58	329
Longitudinal Reinforcement Ratio (LRR) [0.19 - 4.28]	Ratio of steel area in the longitudinal direction to gross section area.	0.0132	303
Transverse Reinforcement Ratio (TRR) [0.06 - 1.73]	Ratio of steel area in the transverse direction to gross section area.	0.0016	266

Table 3.1: Description of target variables, median values, training data sizes

Data cleaning. The data cleaning process, detailed in Section 3.5, was essential for refining the dataset to focus on bridge types most susceptible to shear failure. By systematically removing

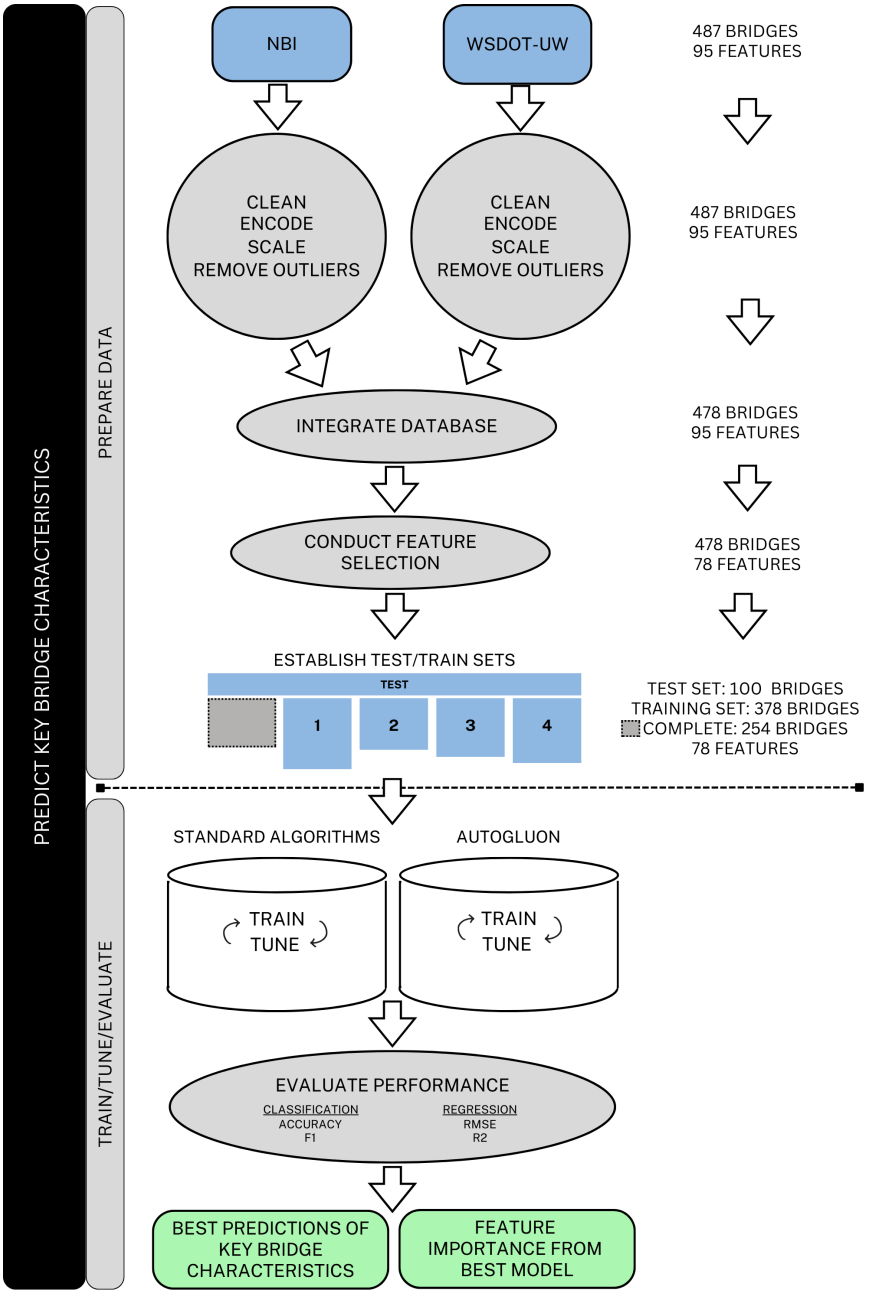


Figure 3.4: Workflow overview for predicting key bridge characteristics. Note: the gray box under the test set represents cases where complete data for all characteristics is available (354 total bridges: 100 in the test set + 254 in the training set). The varying sizes of the blue boxes illustrate differences in the number of training samples available for each target variable, depending on the completeness of the data for that particular characteristic, as shown in Table 3.1.

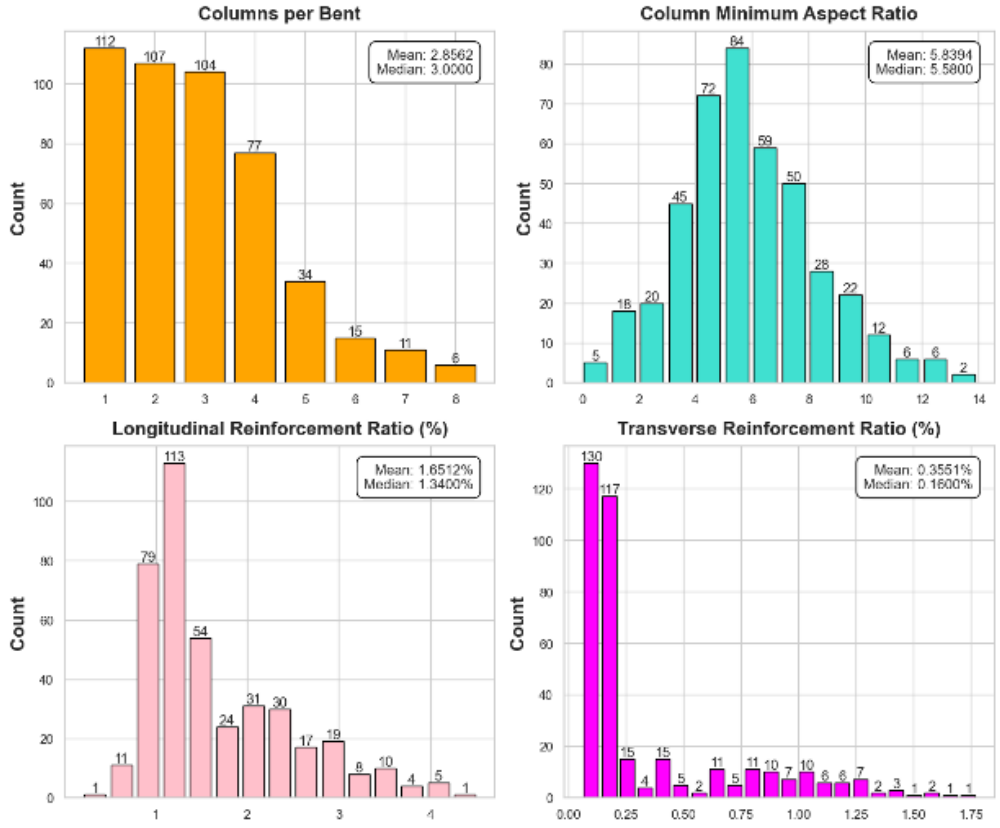


Figure 3.5: Target variable distributions.

samples, we could isolate relevant bridge structures (i.e., multi-span bridges) while eliminating redundant, irrelevant, or incomplete features that could compromise the accuracy and robustness of the modeling process.

Encoding and scaling. Numeric features were scaled using `Scikit-learn`'s `StandardScaler`, standardizing data to a mean of zero and unit variance, essential for algorithms like support vector machines and neural networks. Categorical variables were encoded using `Scikit-learn`'s `OrdinalEncoder`, to maintain consistency in data representation of the NBI while avoiding the dimensionality expansion from one-hot encoding.

Outlier removal. Bridges with target values beyond three standard deviations from the mean were excluded to reduce the influence of atypical observations on model generalizability. Classes with fewer than three instances were removed for classification tasks to address extreme class imbalance.

Feature Selection. Starting with 95 features, we first removed one feature exhibiting low variance (variance threshold < 0.01), as it contributed minimal information to the model. We eliminated sixteen additional features due to high correlation (correlation threshold > 0.90) to mitigate multicollinearity and improve model interpretability. This refinement resulted in a final set of 78 features. See supplemental information in Appendix B for more details.

Integrating Databases and Establishing Training and Testing Sets. We merged the cleaned datasets using a common bridge reference number. Instead of removing bridges with missing data in the target variables, we randomly selected a hold-out test set of 100 bridges with complete data. For each target variable, we created training sets by including all remaining available data.

This process yielded a training dataset representative of the types of bridges without any systematic missing data issue (e.g., due to how the bridge characteristics were recorded). Imputing the missing data can help augment the dataset but introduce modeling bias, thus deserving further attention in future work.

Manual vs. AutoML preprocessing. Although AutoML tools can automate preprocessing tasks, we chose to preprocess the data manually because the relatively small dataset benefited from a more customized approach. The preprocessing steps applied to the data used for AutoML were the same as those used for the standard algorithms, ensuring consistency across all models.

3.6.2 Model training

We trained the models using a diverse set of algorithms described in Section 3.4, including decision trees, random forests, support vector machines, gradient boosting, and neural networks, all implemented via `Scikit-learn`. To ensure robust model performance evaluation, we employed 5-fold cross-validation, splitting the training data into five subsets to evaluate each model iteratively. This approach minimizes overfitting and provides a more reliable estimate of the model's generalization. Hyperparameter tuning was conducted using a randomized search, where a wide range of parameters was explored to optimize model performance based on accuracy for classification tasks and root mean squared error for regression tasks. See supplemental information in Appendix B for more details on hyperparameter ranges.

After implementing the standard algorithms, we further enhanced the analysis by incorporating AutoGluon. AutoGluon automates model selection and hyperparameter optimization by stacking and ensembling models, including those mentioned earlier, to improve predictive performance. All preprocessing steps, including feature scaling, encoding, and outlier handling, remained consistent with those used in the standard algorithms, ensuring a seamless integration with AutoGluon. This preserved the preprocessing pipeline while allowing AutoGluon to efficiently optimize model performance and streamlining the overall workflow.

To ensure that model performance was not overly dependent on a specific test set, we repeated the model training process 10 times, each establishing a new test set by selecting a different random seed. This iterative approach allowed us to evaluate the stability of model performance across different test splits, reducing the likelihood that variations in test set composition would influence the results.

3.6.3 Model evaluation and selection

We assessed performance using the root mean square error (RMSE) and coefficient of determination (R^2) for these regression prediction tasks. RMSE measures the average magnitude of prediction errors, revealing how far predictions deviate from actual values. In contrast, R^2 indicates the proportion of variance in the target variable that the model explains, showing how well it captures underlying data patterns. These metrics provide a balanced view of regression model performance, addressing error magnitude and explanatory power.

To ensure the robustness of our evaluation, we averaged the RMSE and R^2 scores across the 10 iterations described earlier. By repeating the model training and evaluation process with different test sets, we minimized the impact of any single test split on the results. This iterative approach provided a more comprehensive assessment of model performance, accounting for potential variability across different data partitions. The final reported metrics represent the mean values and standard deviations across these iterations, offering a stable and reliable comparison of the predictive capabilities of each model.

After determining the best-performing model based on our established performance metrics, we utilized permutation feature importance to assess the contribution of individual features to the model's predictions. Permutation feature importance is a model-agnostic method that measures the impact of each feature by randomly shuffling its values and observing the corresponding change in the model's performance. A significant drop in performance indicates that the feature is crucial, whereas minimal change suggests a less important feature. By applying this approach, we identified the most influential features driving our predictions, providing insight into the key factors affecting shear-critical bridge identification.

To ensure the stability of the feature importance rankings, we averaged the computed importance scores across the 10 iterations described earlier. This mitigated the impact of variations in test set composition, leading to a more reliable assessment of feature importance. Additionally, we normalized the importance values using Min-Max scaling, which rescales the feature importance scores to a 0–1 range, preserving relative differences while making them easier to compare across the four target variables. This transformation improves interpretability, ensuring that differences in fea-

ture importance across models and targets remain meaningful while preventing scale discrepancies from distorting comparisons.

3.6.4 Results

Table 3.2 presents the performance metrics for the algorithms evaluated across our four target variables: *Columns per Bent*, *Column Minimum Aspect Ratio*, *Longitudinal Reinforcement Ratio*, and *Transverse Reinforcement Ratio*. Performance and their respective standard errors were assessed using root mean squared error (RMSE) and the coefficient of determination (R^2).

For *Columns per Bent*, the AutoML algorithm, AutoGluon, achieved the lowest RMSE (0.945 $\pm$ 0.092) and the highest R^2 score (0.598 $\pm$ 0.035), outperforming the other models. XGBoost (gradient boosting algorithm) and random forest also demonstrated competitive performance with similar RMSE and R^2 values. For *Column Minimum Aspect Ratio*, the random forest algorithm exhibited the lowest RMSE (1.818 $\pm$ 0.144), followed closely by AutoGluon (1.853 $\pm$ 0.149), outperforming the remaining algorithms. The highest R^2 score (0.415 $\pm$ 0.046) was achieved by random forest, indicating a better fit for this target variable. In predicting the *Longitudinal Reinforcement Ratio*, AutoGluon again achieved the lowest RMSE (0.00724 $\pm$ 0.000) and the highest R^2 score (0.154 $\pm$ 0.085). However, all algorithms exhibited relatively low R^2 values, suggesting challenges in accurately modeling this target. For *Transverse Reinforcement Ratio*, random forest obtained the lowest RMSE (0.00173 $\pm$ 0.000) and the highest R^2 score (0.767 $\pm$ 0.069), showing a strong predictive capability. XGBoost and AutoGluon performed similarly, with R^2 scores above 0.75, indicating robust performance for this target.

Figures 3.6 and 3.7 display the performance of six machine learning models across four target variables, based on results from ten iterations. Each figure includes 95% confidence intervals computed using the t-distribution. Figure 3.6 presents Root Mean Squared Error (RMSE), with models ranked from best (lowest RMSE) to worst in each subplot. In contrast, Figure 3.7 reports the corresponding R^2 values, where models are ordered from lowest to highest, with higher R^2 values indicating better predictive performance.

The confidence intervals are interpreted following the principles of Fisher's Least Significant

	Columns per Bent		*Column Minimum Aspect Ratio*		*Longitudinal Reinforcement Ratio*		*Transverse Reinforcement Ratio*	
Model	**RMSE ± SE**	**R^2 ± SE**	**RMSE ± SE**	**R^2 ± SE**	**RMSE ± SE**	**R^2 ± SE**	**RMSE ± SE**	**R^2 ± SE**
Decision Tree	1.285 ± 0.054	0.238 ± 0.069	2.130 ± 0.052	0.194 ± 0.033	0.00789 ± 0.0003	0.000 ± 0.020	0.00248 ± 0.000	0.519 ± 0.044
Random Forest	0.976 ± 0.033	0.570 ± 0.015	**1.818 ± 0.046**	0.415 ± 0.015	0.00730 ± 0.0003	0.138 ± 0.034	**0.00173 ± 0.000**	0.767 ± 0.022
Support Vector Machine	1.100 ± 0.036	0.454 ± 0.019	2.073 ± 0.058	0.240 ± 0.021	0.00782 ± 0.0003	0.019 ± 0.014	0.00241 ± 0.000	0.544 ± 0.037
Gradient Boosting	0.972 ± 0.019	0.573 ± 0.020	1.864 ± 0.021	0.385 ± 0.012	0.00744 ± 0.014	0.108 ± 0.028	0.00174 ± 0.037	0.762 ± 0.020
Neural Network	1.345 ± 0.020	0.181 ± 0.029	2.131 ± 0.012	0.197 ± 0.014	0.00786 ± 0.028	0.007 ± 0.018	0.00221 ± 0.020	0.623 ± 0.020
AutoML	**0.945 ± 0.029**	0.598 ± 0.011	1.853 ± 0.047	0.393 ± 0.016	**0.00724 ± 0.000**	0.154 ± 0.027	0.00175 ± 0.000	0.762 ± 0.019

Table 3.2: Predicting key bridge characteristics model performance metrics (average RMSE and R^2 with standard errors based on 10 iterations)

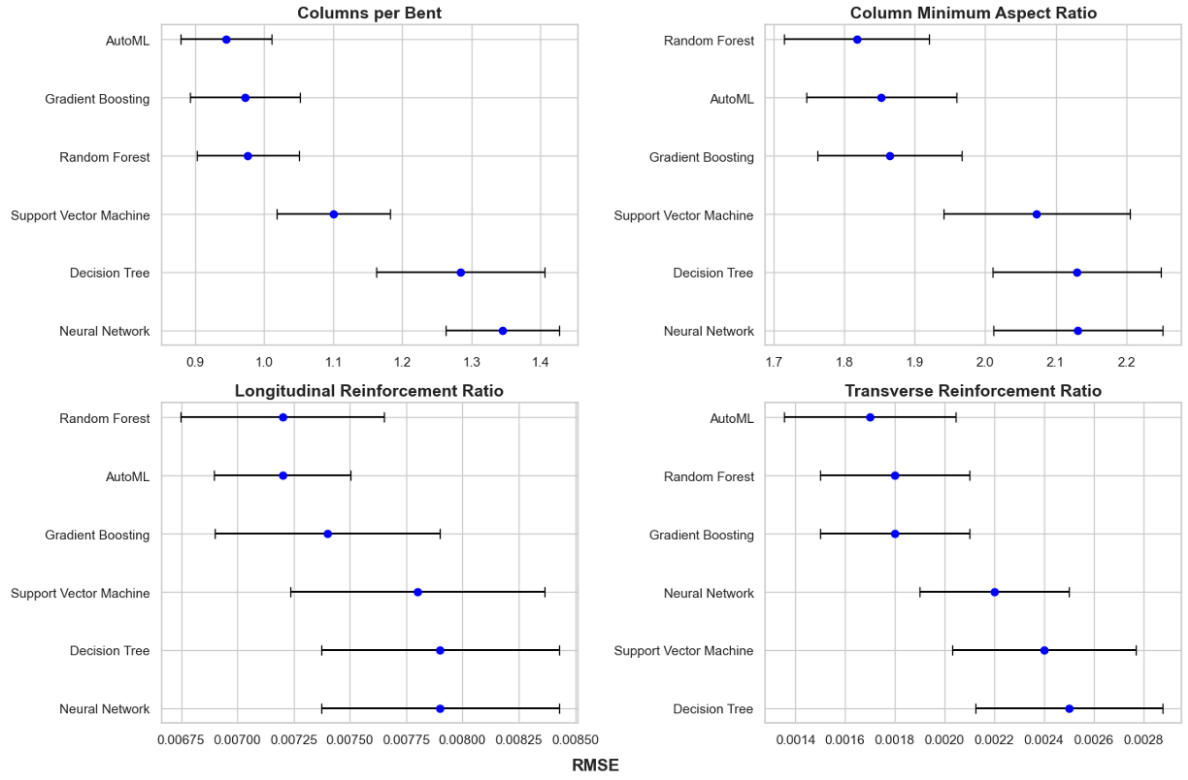


Figure 3.6: Root Mean Squared Error (RMSE) with 95% confidence intervals for six machine learning models across four target variables. Confidence intervals are based on 10 repeated model evaluations and constructed using the t-distribution. Intervals are interpreted in the style of Fisher's Least Significant Difference (LSD): non-overlapping intervals suggest statistically significant differences in performance. Models are ordered from lowest (best) to highest RMSE in each subplot.

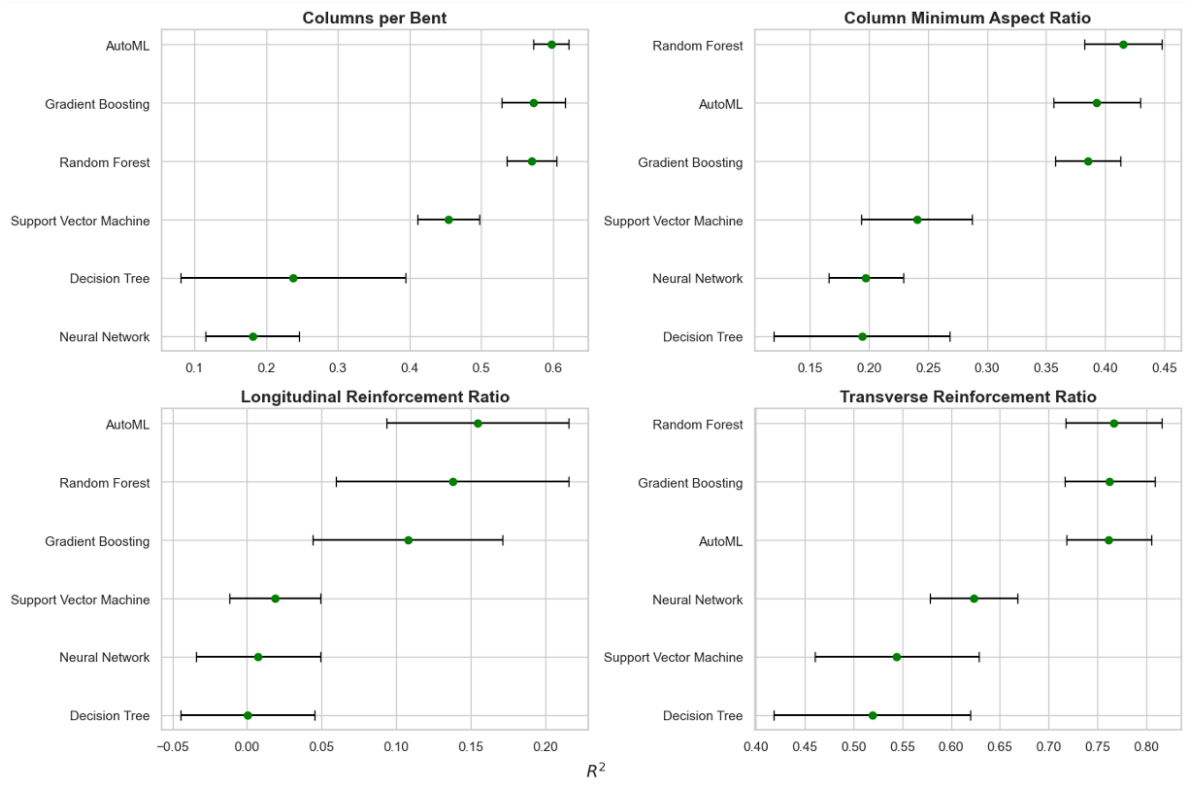


Figure 3.7: R^2 with 95% confidence intervals for six machine learning models across four target variables. Confidence intervals are based on 10 repeated model evaluations and calculated using the t-distribution. Intervals are interpreted in the style of Fisher's Least Significant Difference (LSD): non-overlapping intervals suggest statistically significant differences in model performance. Within each subplot, models are ordered from highest to lowest R^2 to aid comparison, with higher R^2 values indicating better predictive performance.

Difference (LSD)[38], where non-overlapping intervals suggest statistically significant differences in model performance. Across all targets, ensemble-based models, particularly AutoGluon (AutoML algorithm), XGBoost (gradient boosting algorithm), and random forest, consistently outperformed individual learners such as decision tree, support vector machine, and neural network. This is most evident in the prediction of both reinforcement ratios, where AutoML and random forest achieved lower RMSE and higher R^2 scores with relatively tight confidence intervals.

For the two target variables, *Columns per Bent* and *Longitudinal Reinforcement Ratio*, where the AutoML algorithm, AutoGluon, demonstrated the best predictive performance, we analyzed ensemble weight distributions to provide additional insight into the composition of the most effective models. The bar charts shown in Figure 3.8 illustrate the average contributions of individual models to the final ensemble across 10 iterations. In the case of *Columns per Bent*, CatBoost (0.34) and LightGBMXT (0.32) dominate the ensemble, indicating that gradient boosting models played a significant role in capturing the underlying relationships. In contrast, for *Longitudinal Reinforcement Ratio*, the ensemble showed a more balanced contribution among Random Forest, LightGBM, and NeuralNetTorch ($\sim$ 0.14), suggesting that no single model type overwhelmingly influenced the predictions. Meanwhile, a standard Random Forest model yielded the best results for the other two target variables, reinforcing that simpler models can sometimes outperform complex ensembles.

Finally, in Figure 3.9 we present the permutation feature importance results for the four target variables analyzed. Permutation feature importance measures the impact of each feature on model predictions by randomly shuffling its values and observing the resulting drop in model performance. A more considerable drop indicates a more important feature. For *Columns per Bent*, **DECK WIDTH (052)** is the most influential predictor, followed by **YEAR BUILT (027)** and **TOTAL HORIZONTAL CLEARANCE (047)**. In contrast, *Column Minimum Aspect Ratio* is primarily driven by **LENGTH OF STRUCTURE IMPROVEMENT(076)** and **MINIMUM VERTICAL UNDERCLEARANCE (054B).** For the *Longitudinal Reinforcement Ratio*, the importance of features is more evenly distributed, with **YEAR BUILT (027)** and **FUTURE AVERAGE DAILY TRAFFIC (114)** as the main predictors, although other vari-

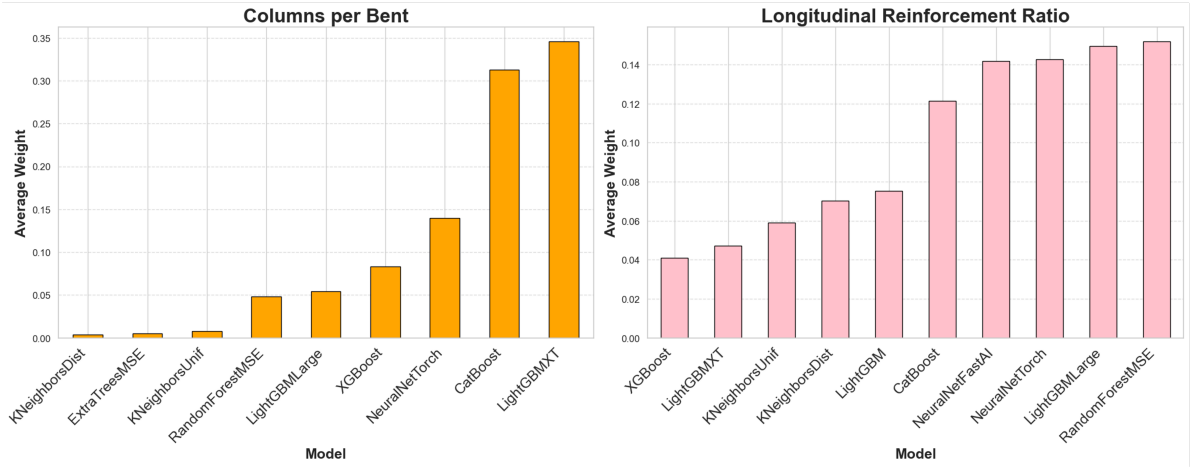


Figure 3.8: Average ensemble weights across iterations for *Columns per Bent* and *Longitudinal Reinforcement Ratio*.

ables also contribute meaningfully. In contrast, *Transverse Reinforcement Ratio* is overwhelmingly dominated by **YEAR BUILT (027)**, with all other features having minimal influence.

Rather than solely reporting predictive performance metrics, we aimed to go beyond the numbers by providing deeper insights into how the models generated predictions. By analyzing ensemble weight distributions and feature importance, we highlighted the key drivers of model behavior, offering a clearer understanding of the decision-making process. This approach enhances the interpretability of the models, allowing for a more informed application in future work.

Further interpretation and implications of these findings are discussed in Section 3.8.

3.7 Applying key bridge characteristic predictions to characterize shear-critical columns

In this section, we evaluate the effectiveness of our predicted bridge characteristics in assessing shear-critical columns and demonstrate how these predictions can enhance regional-scale decision-making. First, we examine how well the predicted characteristics perform when used in calculating and predicting a *Shear-Critical Column Index (SCCI)*, ensuring that the model-derived features provide meaningful insights into structural vulnerability. Then, we showcase the practical application of this index, leveraging the predicted characteristics to prioritize assessments and resource allocation across a network of bridges. Figure 3.10 shows an overview of this framework.

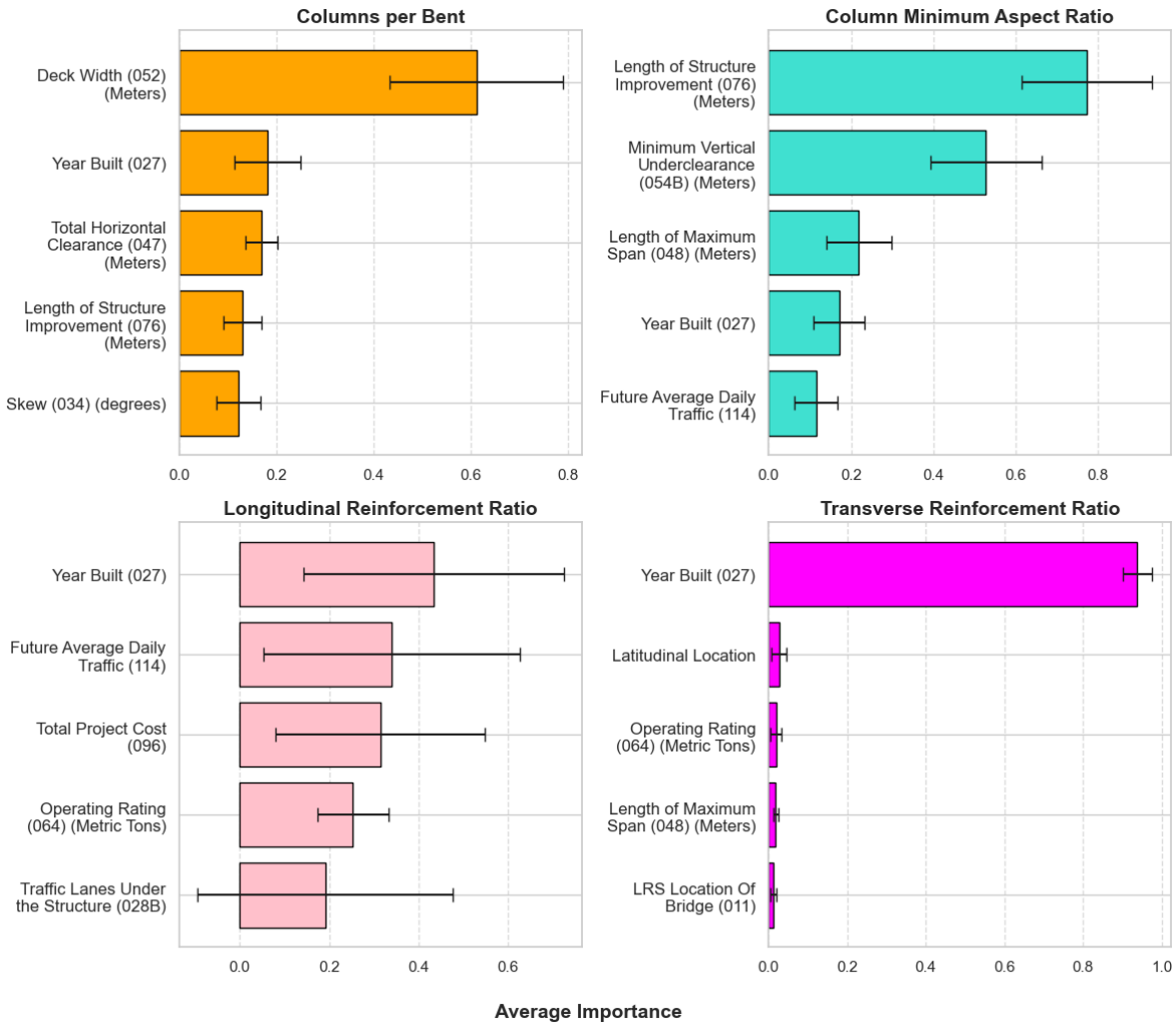


Figure 3.9: Scaled feature importance by target variable. The error bar represents the standard deviation, showing the variability of feature importance across the 10 iterations.

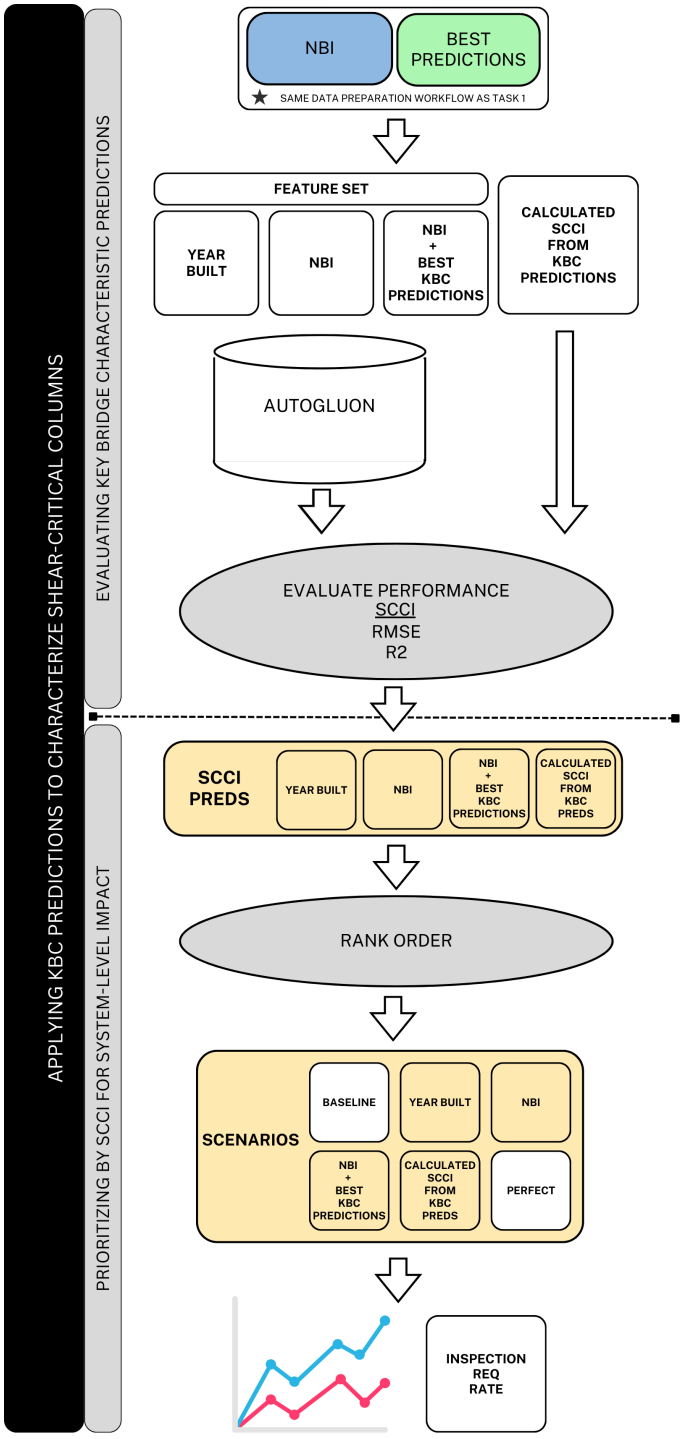


Figure 3.10: Workflow overview of characterizing shear-critical columns.

3.7.1 Shear-Critical Column Index (SCCI)

One of the primary causes of bridge collapse following an earthquake is column shear failure. Shear failure in a column can occur if the shear demand on a column exceeds the shear capacity. During an earthquake, the ground motion can induce large shear forces in columns, and if they are not designed to handle these forces, they may experience shear failure.

One way to understand if a bridge is susceptible to failing in shear is by understanding the relationship between the shear capacity and demand. With the additional key bridge characteristics from the WSDOT/UW Dataset, it is possible to calculate the *Shear-Critical Column Index* given by:

$$SCCI = \frac{\left[1.6\left(\frac{\sqrt{f'_c}}{f_y}\right) + 0.5 \cdot \mathrm{TRR}\left(\frac{f_{yt}}{f_y}\right)\right] \cdot \mathrm{L/H_{min}}}{\left[1.3Z\left(\frac{\pi}{4}\right)\eta \cdot \mathrm{LRR}\right]} \tag{3.1}$$

where the shear capacity of the column over the shear demand on the column is influenced by the compressive strength of concrete, f'_c, the yield strength of longitudinal reinforcement, f_y, the yield strength of transverse reinforcement, f_{yt}, *Transverse Reinforcement Ratio*, TRR, the *Column Minimum Aspect Ratio*, $\mathrm{L/H_{min}}$, a distribution factor based on single vs. multiple columns, Z, the efficiency factor of the column cross-section, η, and the *Longitudinal Reinforcement Ratio*, LRR. Note - our initial analysis used regression to predict the number of columns per pile, as this metric is more relevant to readers. However, for this specific analysis, we only needed to distinguish whether a bent had one or multiple columns. To address this, we applied the same workflow for binary classification to determine Z, with associated results provided in the supplemental information in Appendix B.

Understanding *SCCI* of bridge columns can inform decisions about retrofitting existing structures to enhance resilience, prioritize maintenance efforts, and ensure safety during extreme loading conditions, such as earthquakes. However, as previously discussed, obtaining the necessary inputs for this index requires significant resources and manual data collection efforts. The ability to predict these inputs using limited training data highlights just one example of this approach's broader impact, demonstrating its potential to streamline assessments and support data-driven decision-

making in structural engineering.

3.7.2 Evaluating Key Bridge Characteristic Predictions

To evaluate the effectiveness of our predicted KBCs, we tested four scenarios using machine learning and manual calculations to predict the *SCCI*, as defined in Equation (3.1). For the three machine learning-based scenarios, we followed the same data preparation approach used for KBC predictions, leveraging features available in the NBI. These scenarios differed in their feature sets: (1) Year Built, (2) NBI Only, and (3) NBI + Best KBC Predictions. The Year Built scenario served as a baseline, given that older bridges are generally more vulnerable to seismic hazards [65]. The NBI Only scenario incorporated the same 78 features used for KBC predictions, while the NBI + Best KBC Predictions scenario supplemented NBI features with our best-performing KBC predictions.

For the machine learning models, we employed AutoGluon due to its flexibility, efficiency, and automated hyperparameter tuning across multiple algorithms. Models were trained and optimized using RMSE as the primary loss function.

The fourth scenario took a direct approach, using the KBC predictions as direct inputs to Equation (3.1) to compute the *SCCI* and compare its performance against the machine learning approaches.

We evaluated and compared model performance for all four scenarios using RMSE and R^2, providing a quantitative assessment of predictive accuracy across different modeling approaches.

3.7.3 Prioritizing by SCCI for system-level impact

Next, we leveraged the power of the extended inventory by looking holistically at the system-level impact by prioritizing columns at the highest risk of shear failure based on their *SCCI* and evaluating the change in inspection rate.

We begin by ranking the bridges according to their *SCCI*, where lower values signify a greater likelihood of shear failure. Once the rank orders are established, we evaluate the effectiveness of our model using specific metrics. We calculate the cumulative number of correctly identified

shear-critical bridges in groups of 10 to make the results easier to interpret. This grouping allows for a clearer understanding of the identification progress as more bridges are evaluated. We then compare this cumulative identification curve against a random inspection and a perfect knowledge scenario.

The random inspection scenario reflects the cumulative percentage of positively identified shear-critical bridges based on the probability of a bridge being shear-critical within each group of 10. For example, if 21 out of 100 bridges are shear-critical, then, on average, 21% of the bridges inspected in each group of 10 would be correctly identified as shear-critical. The random inspection scenario on a graph would appear as a steadily rising line with a slope corresponding to the proportion of shear-critical bridges. This scenario serves as a baseline for comparison, showing the expected identification rate if bridges were randomly selected for inspection.

While unrealistic in practice, the perfect knowledge scenario represents the ideal case where we have complete and accurate information about each bridge's *SCCI* magnitude, allowing us to prioritize inspecting all shear-critical bridges first. In this situation, we could identify every shear-critical bridge before any non-shear-critical ones, resulting in a cumulative identification curve that rises sharply and reaches the maximum number of shear-critical bridges early. This scenario is depicted on a graph as a steep, linear ascent until all shear-critical bridges are identified, followed by a horizontal line, indicating that the remaining inspections only involve non-shear-critical bridges.

These two benchmarks provide reference points to assess the performance of our model. We can evaluate how well the model prioritizes high-risk bridges by comparing the model's cumulative identification curve against the random inspection and perfect knowledge scenarios. The closer the model's curve aligns with the perfect knowledge scenario, the more effective it is at identifying shear-critical bridges early. In contrast, a performance closer to the random scenario suggests slight improvement over random selection. This comparison helps determine the model's value in prioritizing inspections based on predicted *SCCI*.

3.7.4 Results

The results from our four scenarios are shown in Table 3.3 and indicate that the Predicted *SCCI* from NBI + KBC Predictions achieved the best performance, with the lowest RMSE (0.928 $\pm$ 0.031) and highest R^2 (0.174 $\pm$ 0.050), demonstrating the effectiveness of using predicted key bridge characteristics as additional features when predicting the *SCCI*. The calculated *SCCI* from KBC predictions also produced competitive results, though with a slightly higher RMSE and lower R^2. In contrast, the NBI Only model exhibited the highest RMSE and lowest R^2, indicating that it was the least effective predictor of *SCCI*.

It is important to note that these results represent averages over 10 iterations, with standard errors reported in Table 3.3 to account for variability across different test sets. This iterative approach ensures that the reported performance metrics are robust and not unduly influenced by a single data partition.

Figures 3.11 and 3.12 show the RMSE and R^2 performance, respectively, of four predictive scenarios evaluated over 10 iterations. Confidence intervals represent 95% confidence bounds based on the t-distribution and are interpreted using Fisher's Least Significant Difference (LSD) framework where non-overlapping intervals suggest statistically significant differences. While models incorporating inferred key bridge characteristics (KBC), either directly or through the SCCI, exhibited slightly better average RMSE and R^2 values compared to baselines relying solely on Year Built or NBI Only, the overlapping confidence intervals indicate that these differences are not statistically significant. Nonetheless, the trend suggests a potential performance benefit from integrating structural inference features, warranting further exploration.

In our prioritization analysis, we evaluated the effectiveness of the NBI + Best KBC Predictions and Calculated SCCI from predicted key bridge characteristics scenarios, comparing them to the random inspection and perfect knowledge reference scenarios. Figure 3.13 illustrates the cumulative identification curve. Notably, both approaches show marked improvement over the Random Inspection scenario, demonstrating a clear enhancement in prioritization effectiveness.

A key point in the curve occurs at 60 bridges inspected, where both the NBI + KBC Predictions and Calculated SCCI approach correctly identify over 80% of shear-critical bridges,

Scenario	Prediction of *SCCI*	
	RMSE ± SE	**R^2 ± SE**
Year Built	0.956 ± 0.031	0.131 ± 0.033
NBI Only	0.961 ± 0.035	0.119 ± 0.051
NBI + Best KBC Predictions	**0.928 ± 0.031**	**0.174 ± 0.050**
Calculated SCCI from KBC Predictions	0.948 ± 0.018	0.133 ± 0.056

Table 3.3: Characterizing shear-critical columns model performance metrics (average RMSE and R^2 with standard errors based on 10 iterations).

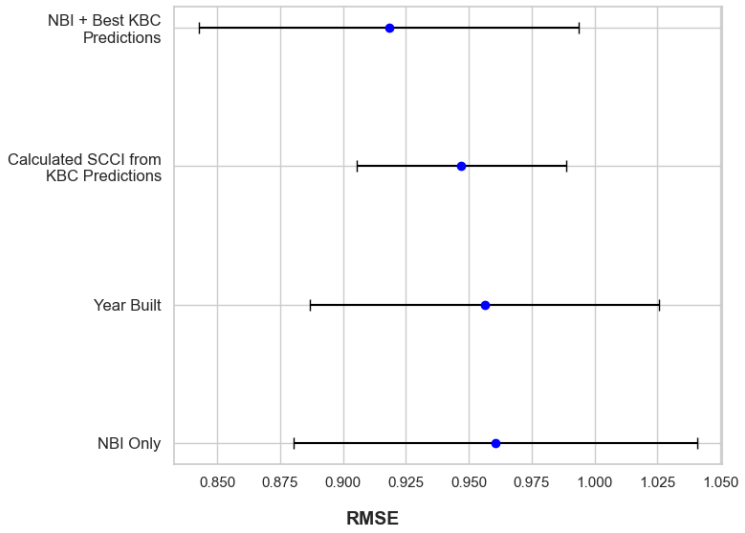


Figure 3.11: RMSE with 95% confidence intervals for prediction of Shear-Critical Column Index (SCCI) for four scenarios. Confidence intervals are based on 10 repeated model evaluations and calculated using the t-distribution. Intervals are interpreted in the style of Fisher's Least Significant Difference (LSD): non-overlapping intervals suggest statistically significant differences in model performance. Scenarios are ordered from lowest to highest RMSE to aid comparison, with lowest RMSE values indicating better predictive performance.

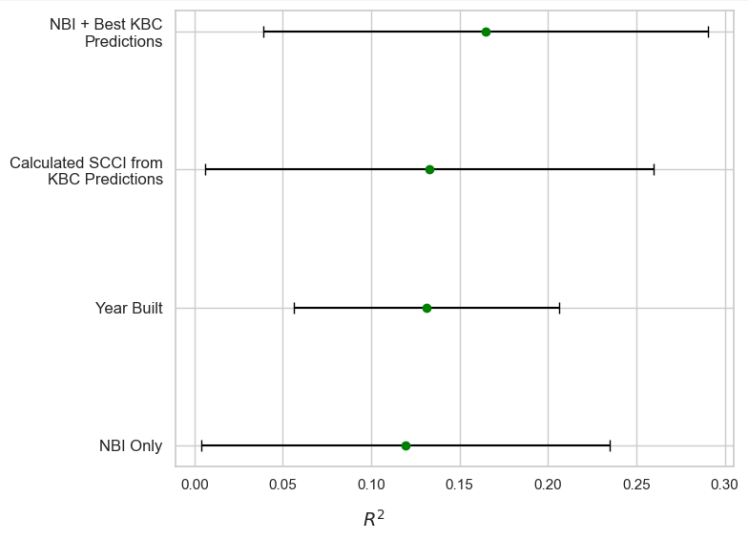


Figure 3.12: R^2 with 95% confidence intervals for prediction of Shear-Critical Column Index (SCCI) for four scenarios. Confidence intervals are based on 10 repeated model evaluations and calculated using the t-distribution. Intervals are interpreted in the style of Fisher's Least Significant Difference (LSD): non-overlapping intervals suggest statistically significant differences in model performance. Scenarios are ordered from highest to lowest R^2 to aid comparison, with higher R^2 values indicating better predictive performance.

significantly outperforming Random Inspection, which identifies only 60% at the same inspection level.

3.8 Discussion

3.8.1 Interpreting the results

In Section 3.6.4, we presented our predictive modeling results, weighted ensemble averages, and feature importances for four target features. As noted, Random Forest provided the best predictive performance for *Column Minimum Aspect Ratio* and *Transverse Reinforcement Ratio* and was also identified as the top weighted algorithm for Autogluon's ensembling methods for *Longitudinal Reinforcement Ratio*. This strong performance can be attributed to Random Forest's ability to handle complex, nonlinear relationships and interactions among features while maintaining robustness against overfitting. Given the heterogeneity of bridge characteristics, Random Forest's ensemble approach—aggregating multiple decision trees—effectively captures underlying patterns

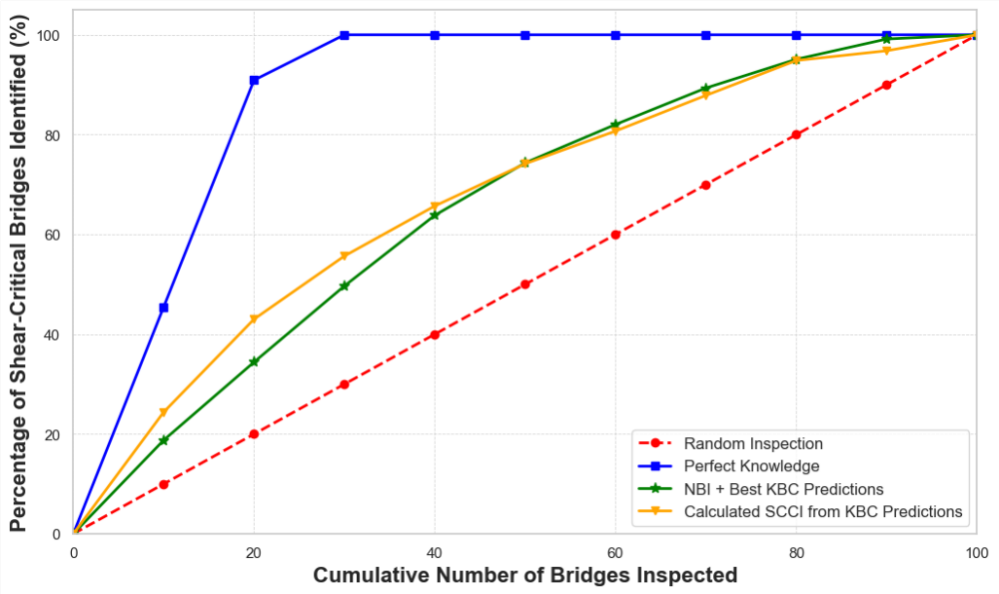


Figure 3.13: Percent of shear-critical bridges identified by shear-critical bridges inspected when rank ordering based on predicted/calculated *SCCI*.

while reducing variance. Additionally, its ability to naturally handle mixed data types makes it well-suited for the National Bridge Inventory dataset. The feature importance rankings reinforce this, as Random Forest leverages the most relevant predictors without relying heavily on strict parametric assumptions.

Notably, boosting algorithms (CatBoost, LightGBMXT, and XGBoost) performed well in the weighted ensembles for *Columns per Bent*. Boosting methods excel in capturing complex, nonlinear relationships and feature interactions that traditional tree-based models may not readily identify. Their iterative learning process allows them to correct errors at each step, enhancing predictive accuracy. Additionally, their ability to prioritize important features without overfitting makes them well-suited for structured datasets like the National Bridge Inventory, where certain bridge characteristics can have subtle but significant influences on structural performance.

The feature importance results provide a mix of expected confirmations and surprising insights, offering a deeper understanding of the model's decision-making process. For example, **YEAR BUILT (027)** ranked among the top five features for every target, reinforcing its strong correlation

with bridge vulnerability and seismic performance. Similarly, **DECK WIDTH (052)** emerged as a key predictor for *Columns per Bent*, aligning with the expectation that wider bridges typically require more support columns. These reaffirming results increase confidence that the model is identifying structurally meaningful relationships. However, other findings warrant further exploration, such as the influence of **TOTAL PROJECT COST (096)** on *Longitudinal Reinforcement Ratio* and the role of **LRS LOCATION OF BRIDGE (011)** in *Transverse Reinforcement Ratio* predictions. These unexpected patterns suggest potential hidden dependencies that could be investigated further. Beyond interpretability, feature importance can also guide feature selection and engineering to improve model performance, helping refine predictive models for future structural assessments.

In Table 3.3, we demonstrate that NBI + Best KBC Predictions outperforms both the Year Built and NBI Only baselines in terms of RMSE. While the improvement is marginal, it highlights the potential of incorporating predicted Key Bridge Characteristics (KBCs) to enhance predictive accuracy. This serves as a promising step, and with continued refinements in modeling techniques and data collection efforts, we can further improve performance in future work. Similarly, Figure 9 illustrates that NBI + Best KBC Predictions achieves better prioritization than random inspection, which is particularly meaningful given the low occurrence rate of shear-critical bridges ($\sim$25%). Despite this needle-in-a-haystack challenge, the ability to inspect just 60% of bridges while identifying 80% of shear-critical structures represents a significant efficiency gain. This approach could translate to substantial cost and resource savings while improving seismic risk mitigation strategies when scaled to a statewide or national level.

3.8.2 Practical insights for using the National Bridge Inventory in machine learning

One of the most important takeaways from our work with the National Bridge Inventory (NBI) dataset is that simpler models performed better than complex ones. While the allure of deep neural networks and extensive AutoML tuning is strong, our experiments consistently showed that reducing complexity often led to better results. We tested deep neural networks with multiple layers, lengthy training times for AutoML algorithms, and models incorporating as many features

as possible. However, in nearly every case, simpler approaches outperformed more complex ones. Shallow neural networks provided better predictive accuracy than deep architectures, likely due to the relatively small dataset size, the relatively structured nature of the dataset, and the risk of overfitting with deeper models. Similarly, AutoGluon produced more effective models when training times were limited, suggesting that prolonged optimization did not necessarily yield better generalization. Additionally, reducing the feature set, rather than including every possible variable, led to improved model performance, reinforcing that feature selection is critical when working with the NBI dataset. These findings suggest that future researchers using the NBI dataset should prioritize interpretability and efficiency. Rather than defaulting to highly complex models, an iterative approach that starts with simpler models and gradually increases complexity (if necessary) will likely yield the best results.

Additionally, while we found AutoGluon to be an effective and efficient AutoML framework, its most significant value was as a tool for guiding model selection rather than serving as the final solution. AutoGluon excels at identifying promising algorithms and constructing ensemble models, but our results showed that its default "best" ensemble was not always the most optimal choice. In many cases, individual models identified by AutoGluon performed just as well or better, suggesting that manual evaluation of its outputs can further refine predictive performance. Additionally, we found that preprocessing the data ourselves before feeding it into AutoGluon produced superior results compared to relying solely on AutoGluon's built-in preprocessing. By handling feature selection, scaling, and encoding externally, we had greater control over data consistency, which improved model performance. These findings suggest that while AutoML can significantly streamline the model selection process, researchers working with the NBI dataset should remain actively involved in preprocessing decisions and model evaluation rather than relying entirely on automated pipelines. A hybrid approach—using AutoML for initial insights but applying domain expertise for refinement—proved the most effective strategy.

While the National Bridge Inventory (NBI) has proven valuable for predictive modeling, certain target variables may not be well-suited for this dataset alone. Structural characteristics and bridge performance involve complex, multi-faceted factors that the available NBI attributes may not fully

capture. However, the rapid growth of machine learning and artificial intelligence presents exciting opportunities to integrate diverse data sources—including imaging, remote sensing, and text-based embeddings—to enhance predictive accuracy. Expanding the available inventory of data sources will be critical for moving toward system-level modeling, where bridges are assessed not just as isolated structures but within the broader transportation network. Combining structured datasets like the NBI with modern AI advancements opens the door for more comprehensive modeling, improving infrastructure assessment and decision-making. As these tools evolve, so will the potential for leveraging data-driven insights to streamline bridge evaluations, optimize resource allocation, and enhance structural resilience.

Chapter 4

INTEGRATING NATURAL LANGUAGE PROCESSING AND UNCERTAINTY QUANTIFICATION FOR DATA-EFFICIENT SEISMIC VULNERABILITY ASSESSMENT

4.1 Abstract

Machine learning offers powerful tools for advancing critical infrastructure seismic vulnerability assessment; however, its effectiveness depends heavily on available data quality, quantity, and structure. Data collection, curation, and integration processes introduce unique challenges, often resulting in high-dimensional datasets with limited sample sizes. This chapter presents two complementary frameworks designed to address these limitations by improving data efficiency and interoperability in seismic risk modeling. The first framework explores the use of pretrained `Word2Vec` embeddings to transform categorical bridge attributes into dense vector representations, demonstrating the potential to improve classification and regression performance compared to traditional ordinal encoding. The second framework introduces conformal prediction to quantify model uncertainty across classification and regression tasks. When applied to key bridge characteristics and the *Shear-Critical Column Index (SCCI)*, this method produces calibrated prediction sets and intervals with coverage near the 90% confidence level. Together, these contributions advance a scalable, interpretable, and data-efficient approach to seismic vulnerability modeling, balancing automation with actionable insights to support more informed infrastructure decision-making.

4.2 Introduction

The adoption of machine learning and artificial intelligence in seismic risk assessment is increasing; however, their effectiveness depends on the availability, quantity, and quality of data, a well-documented challenge in the field [63]. This challenge is particularly evident when assessing bridge and bridge system fragility, where diverse data sources, such as structural design specifications,

inspection reports, and material properties, often come in various formats, making integration into a machine-readable format difficult [80, 114, 20]. We encountered this same challenge in Chapter 3, where limited data availability and the labor-intensive process of manually extracting measurements from bridge blueprints constrained our usable dataset to just 400 bridges. While the National Bridge Inventory (NBI) provides a rich source of information, it does not always include the specific target variables most relevant to seismic vulnerability. This was particularly true for critical indicators such as *Column Minimum Aspect Ratio* and *Longitudinal Reinforcement Ratio*, which were central to our analysis in Chapter 3. The mismatch between available features and desired targets produced a dataset with high dimensionality but a relatively small sample size, complicating efforts to develop robust and efficient predictive models.

To address these limitations, we present two complementary frameworks that streamline data processing and enhance its utility, enabling more effective analysis and interpretation. First, we explore the use of text through text embedding techniques in natural language processing (NLP) to utilize categorical variables better, transforming textual data into a more informative and structured format. This approach enhances both data quantity and quality, improving model performance. Second, we integrate uncertainty quantification (UQ) to account for model uncertainties, ensuring more reliable interpretations, particularly in data-limited scenarios. By combining NLP-driven data enhancement with UQ, we aim to provide more data-efficient approaches for analyzing shear risks in bridge columns, leading to more informed decision-making.

4.3 Natural Language Processing of categorical variables

4.3.1 Background

In Chapter 3, we utilized the National Bridge Inventory (NBI) to predict Key Bridge Characteristics (KBCs) associated with the seismic vulnerability of bridges. Among the 95 available features in the inventory, 50 were categorical variables containing text data. Traditional machine learning approaches require categorical variables to be represented numerically through encoding techniques, such as one-hot or label encoding, which transform categorical values into numerical representations that machine learning algorithms can process. However, these encodings do not always accurately

capture the meaning or context of the categories, which can diminish the model's predictive performance.

Recent Natural Language Processing (NLP) advancements provide a powerful alternative to traditional encoding methods. As a subfield of artificial intelligence (AI), NLP enables computers to understand, interpret, and generate human language, making it particularly effective for extracting meaningful representations from textual data. In machine learning, word embeddings offer a more sophisticated approach than conventional encodings. Unlike one-hot encoding or bag-of-words representations, which treat words as discrete entities without capturing their relationships, word embeddings—such as `Word2Vec`[68]and `FastText`[13]—represent words as dense vectors in a continuous space. These embeddings preserve syntactic and semantic similarities, allowing models to recognize contextual meanings and linguistic nuances. By transforming textual information into numerical vectors, NLP techniques enable machine learning models to identify patterns, uncover relationships, and extract deeper insights that traditional encoding methods often fail to capture.

In this work, we introduce a framework for implementing word embeddings of NBI categorical variables using `Word2Vec` and enhancing their interpretability through t-distributed Stochastic Neighbor Embedding (t-SNE) [101] for dimensionality reduction. We apply this approach to a newly introduced multiclass target, first implementing our predictive workflow from Chapter 3 to extract key bridge characteristics and then evaluating the predictive performance of this embedding method against standard ordinal encoding (employed in Chapter 3). Our results demonstrate the potential of word embeddings to improve feature representation and enhance machine learning performance on NBI data.

4.3.2 Method

One of the key contributions of Chapter 3 is the development of an approach to expand the inventory of bridge characteristics relevant to seismic vulnerability. To further demonstrate the versatility and extensibility of this approach, we introduce an additional target: *Object Intersected*. This feature categorizes the type of geographical element a bridge crosses, such as an overpass, waterway, or ramp. Including this target serves multiple purposes. First, it illustrates that the framework is not

limited to a handpicked set of characteristics but is generalizable to other attributes of interest, particularly those identified by stakeholders as valuable for planning and hazard mitigation. Second, it helps lay the groundwork for Chapter 4, where we demonstrate how the proposed NLP framework can complement existing machine learning workflows and enhance downstream analysis.

We employ our machine learning framework from Chapter 3 to identify the optimal algorithm for predicting this characteristic, assessing model performance based on the overall accuracy of this multiclass target. To enhance predictive power, we leverage permutation feature importance to identify the most influential features. We then implement `Word2Vec` embeddings, apply dimensionality reduction with t-SNE, and compare the results to ordinal encoding, evaluating which approach yields the best predictive performance. Figure 4.1 shows an overview of this approach.

`Word2Vec`[68] is a neural network-based technique that transforms words into continuous vector representations, capturing semantic relationships between words based on their contextual similarity. Unlike traditional encoding methods, `Word2Vec` generates dense, word vectors that preserve meaningful linguistic patterns, making it well-suited for machine learning tasks. However, due to the high dimensionality of word embeddings and the small size of our dataset, we applied t-Distributed Stochastic Neighbor Embedding (t-SNE)[101] for dimensionality reduction. t-SNE is a non-linear technique that maps high-dimensional data into a lower-dimensional space while preserving local structure, helping mitigate the risk of overfitting and improving model performance by reducing noise in the feature space. Through an iterative process, we determined the most suitable combination of embedding technique and dimensionality for t-SNE, ensuring that the reduced representations effectively captured the underlying structure of our data. Given our dataset's small size and high proportion of categorical variables, this approach proved more suitable than alternatives like `FastText` or `BERT`[26], which rely on large corpora to generate robust embeddings. `FastText`'s subword information and `BERT`'s deep contextual representations, while powerful, introduce complexities and computational demands that are less effective in our setting. The t-SNE approach, coupled with `Word2Vec` embeddings, allowed us to preserve meaningful relationships while mitigating the risk of overfitting.

We use the best-performing algorithm and the same hyperparameter tuning approach from the

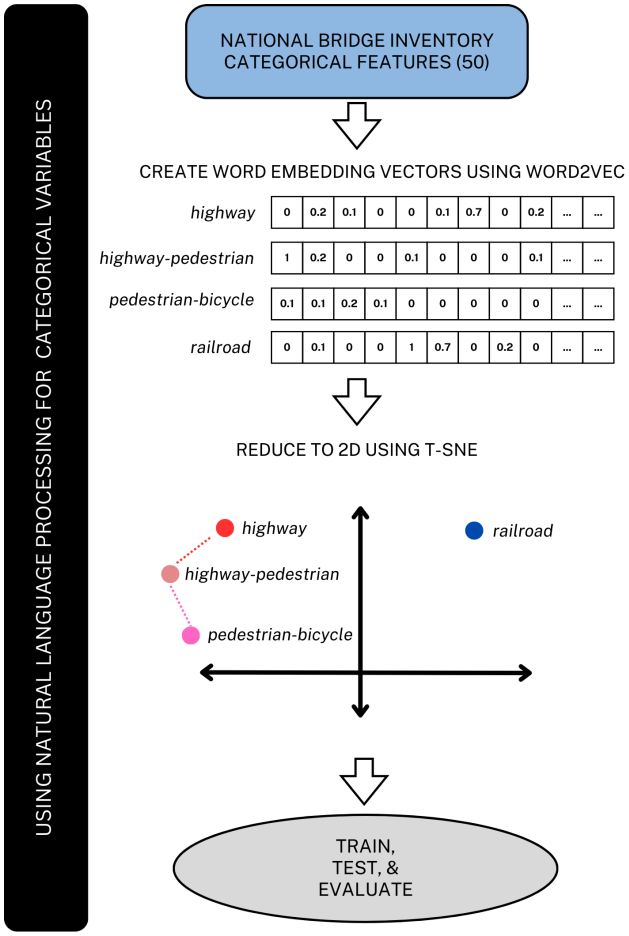


Figure 4.1: Workflow overview for word embeddings

original modeling effort in Chapter 3 to compare performance between models based on Accuracy and F1 score for classification tasks and RMSE and R^2 for regression tasks. Though `Word2Vec`was best suited for our data set, additional results and analysis of alternative embedding methods that we implemented can be found in the supplemental information in Appendix C.

4.3.3 Results

We began our analysis using our machine learning framework from Chapter 3 to determine the best algorithm and hyperparameters for predicting *Object Intersected*. As part of our exploratory data analysis, Figure 4.2 shows the distribution of geographical features/objects that each bridge in the dataset intersects. The data set comprises 779 entries across six classes: overpass, water, ramp, underpass, terrain, and viaduct. Viaduct and terrain classes have only ten and nine samples, respectively. This class imbalance can impact model performance, particularly in classification tasks, as models may become biased toward the majority classes and struggle to learn meaningful patterns from underrepresented ones. As a result, predictions for minority classes may be less reliable, and evaluation metrics such as overall accuracy could be misleading. We used a secondary metric, the F1 score, which balances precision and recall and provides a more meaningful evaluation for imbalanced classification problems to address this limitation.

Table 4.1 presents the evaluated algorithms' predictive performance, including decision tree, random forest, support vector machine, gradient boosting, neural network, and the AutoML framework AutoGluon. Among these, gradient boosting (specifically, XGBoost) achieved the highest performance, outperforming all other models in both accuracy and F1 score. Random forest and AutoGluon also demonstrated strong results, closely trailing XGBoost. Support vector machine and decision tree produced competitive outcomes, though they fell slightly short of the top-performing models. Neural networks, while effective in some cases, showed the greatest variability in performance, as evidenced by a more significant standard error. Performance metrics were averaged across ten random iterations to ensure robust evaluation and mitigate the influence of any single training-test split. Further details on the hyperparameter tuning process are provided in the supplemental information in Appendix C.

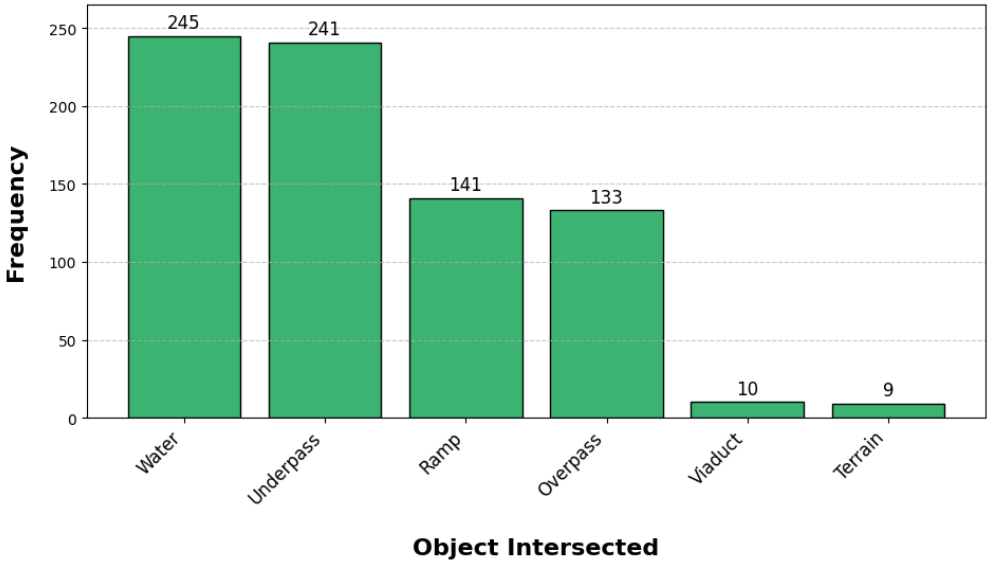


Figure 4.2: Histogram of *Object Intersected*

Model	**Accuracy ± SE**	**F1 score ± SE**
Decision Tree	86.8% ± 1.4%	86.6% ± 1.4%
Random Forest	89.6% ± 1.0%	89.1% ± 1.0%
Support Vector Machine	87.5% ± 0.9%	86.3% ± 0.9%
Gradient Boosting	**90.4% ± 0.9%**	**89.8% ± 1.2%**
Neural Network	85.6% ± 1.2%	84.6% ± 1.7%
AutoML	89.9% ± 0.8%	88.8% ± 0.9%

Table 4.1: Predicting *Object Intersected* model performance metrics (average accuracy and F1 score with standard errors for *Object Intersected* across 10 iterations

Since gradient boosting yielded the best overall performance, we selected XGBoost as the benchmark model for comparing our encoding and embedding approaches. Figure 4.3 highlights the top eleven features, as determined by permutation feature importance derived from the XGBoost model. We chose the top eleven features because they represented a natural inflection point in importance scores—beyond the eleventh feature, importance values dropped off significantly (see Figure C.1 in supplemental information in Appendix C). Among these, the **bolded** features are categorical variables, corresponding to the mappings provided in Table 4.2. To enhance these categorical features, we embedded the mapped values using `Word2Vec`, replacing simple categorical codes with dense vector representations that better captured the contextual relationships between categories.

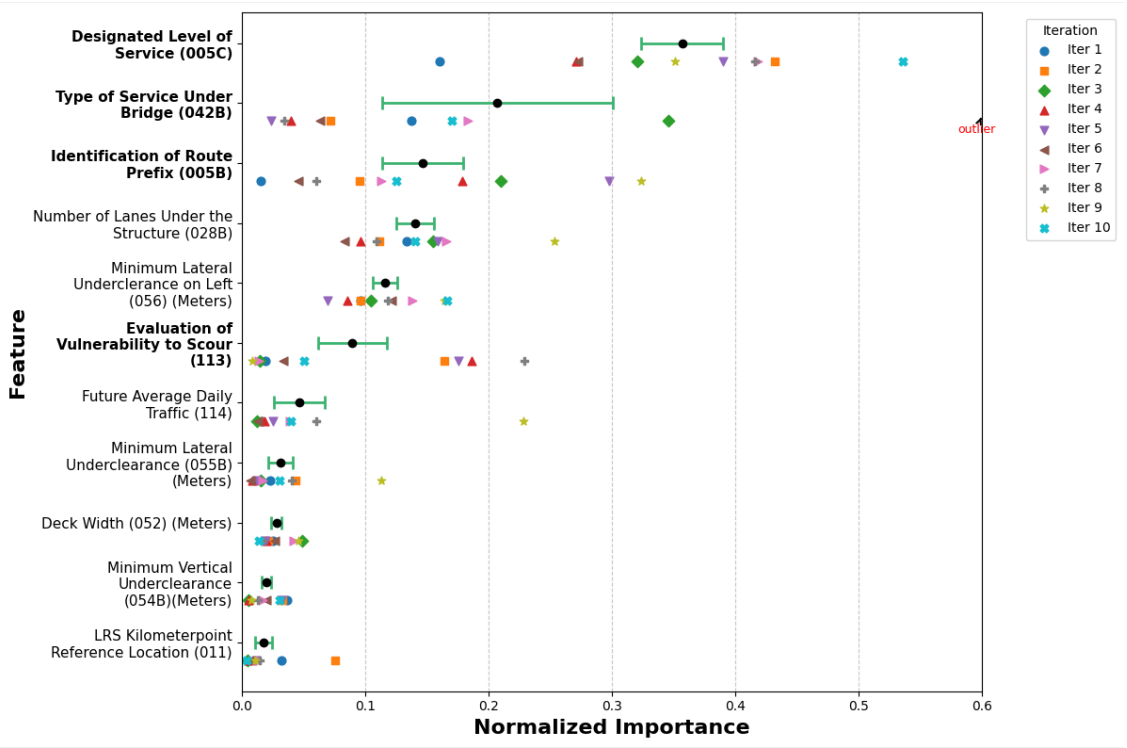


Figure 4.3: Mean scaled feature importance ± standard error for modeling of *Object Intersected*. **Bolded** features are categorical variables.

We used a pre-trained version of `Word2Vec` (trained on Google News 300) to embed four categorical features—**SERVICE LEVEL 005C**, **ROUTE PREFIX 005B**, **SERVICE UNDER**

042B, and **SCOUR CRITICAL 113**—into 300-dimensional vectors. While this approach enriched the feature representations, it also introduced substantial dimensionality, posing risks of overfitting and increased computational cost. To address this, we applied t-SNE to reduce each categorical feature to two dimensions. The other seven non-categorical features were retained in their original form and included alongside the two-dimensional embeddings during model training and testing. This approach allowed us to preserve essential structure from the embeddings while maintaining a compact and manageable feature set compatible with gradient boosting models. Figures 4.4, 4.5, 4.6, and 4.7 visualize the t-SNE 2D `Word2Vec` embeddings of the four categorical features.

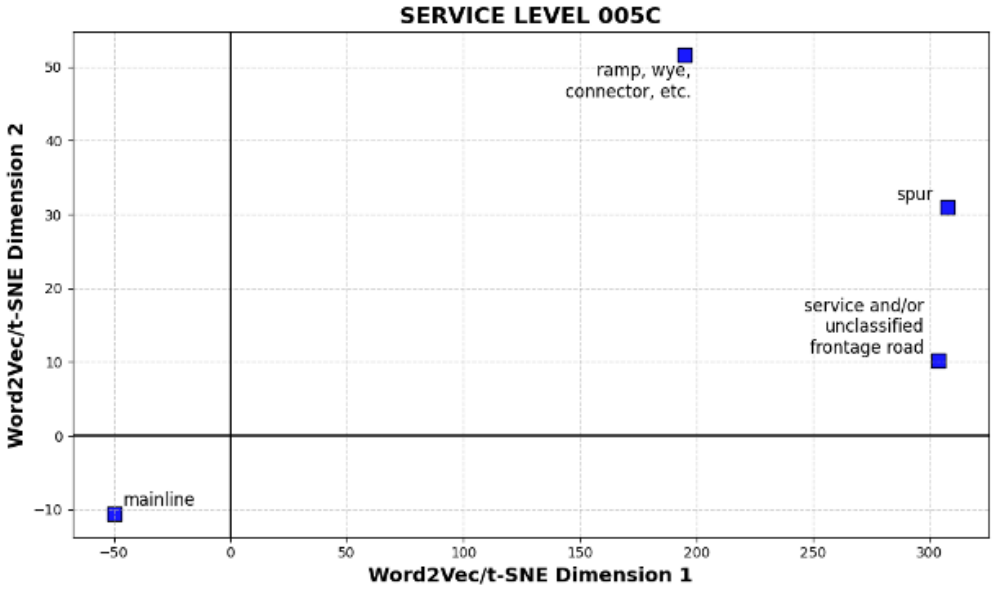


Figure 4.4: Visualization of embeddings for **SERVICE LEVEL 005C**

The `Word2Vec`/t-SNE visualizations reveal clear semantic distinctions across the categorical features. For **SERVICE UND 042B** (Figure 4.5), categories such as "highway-waterway-railroad" and "pedestrian-bicycle" cluster together, indicating shared contextual usage, while "waterway", "highway with or without pedestrian", and "other" appear more isolated. In **SCOUR CRITICAL 113** (Figure 4.7), "bridge foundations assessed as stable" group closely, whereas bridges

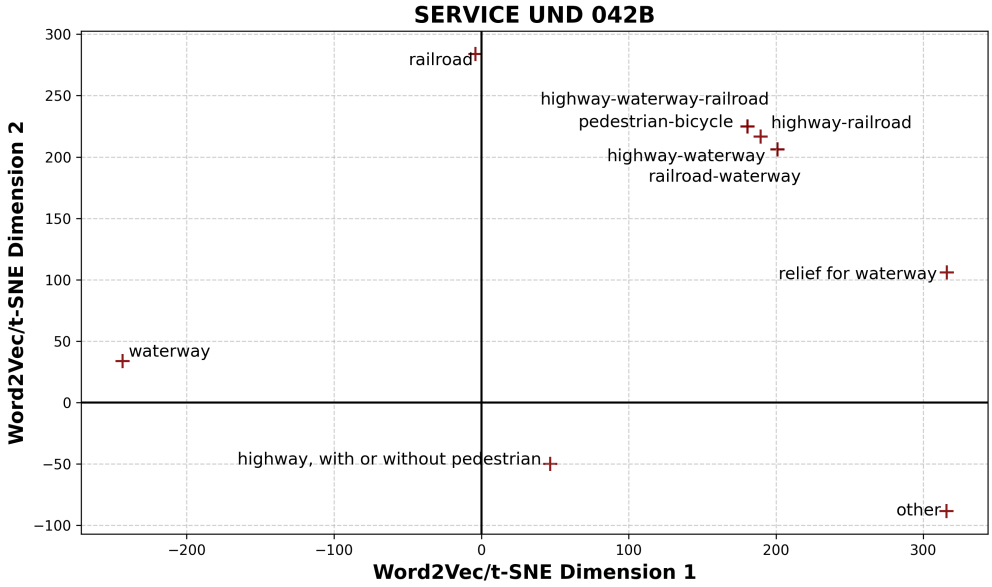


Figure 4.5: Visualization of embeddings for **SERVICE UND 042B**

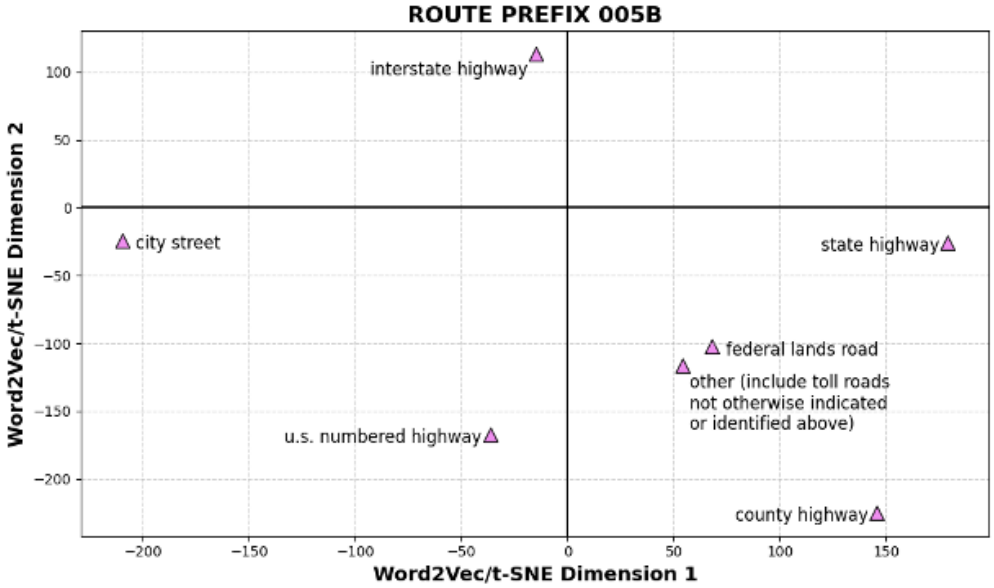


Figure 4.6: Visualization of embeddings for **ROUTE PREFIX 005B**

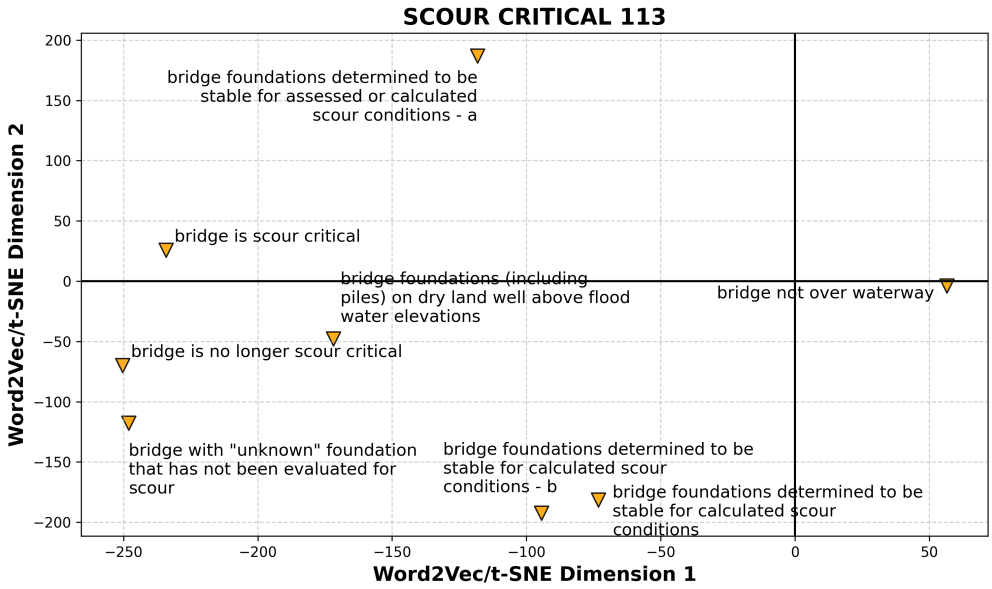


Figure 4.7: Visualization of embeddings for **SCOUR CRITICAL 113**

classified as "scour critical" or "with unknown foundations" are spatially distinct, highlighting risk differentiation. Similarly, the embedding for **ROUTE PREFIX 005B** (Figure 4.6) shows that "city street" and "U.S. numbered highway" are well separated from secondary and administrative road types, such as "state highway" and "federal lands road", which cluster together. Lastly, the **SERVICE LEVEL 005C** embedding (Figure 4.4) reveals "mainline" as distinct from secondary service types like "ramp, wye, connector" and "spur", suggesting functional separation between primary and auxiliary routes. Together, these visualizations demonstrate the `Word2Vec` model's ability to capture nuanced contextual relationships among infrastructure categories.

Code	Description
	SERVICE LEVEL 005C
0	None of the below
1	Mainline
2	Alternate
3	Bypass

4	Spur
6	Business
7	Ramp, Wye, Connector, etc.
8	Service and/or unclassified frontage road

SERVICE UND 042B

1	Highway, with or without predestrian
2	Railroad
3	Pedestrian-bicycle
4	Highway-railroad
5	Waterway
6	Highway-waterway
7	Railroad-waterway
8	Highway-waterway-railroad
9	Relief for waterway
0	Other

ROUTE PREFIX 005B

1	Interstate highway
2	U.S. numbered highway
3	State highway
4	County highway
5	City street
6	Federal lands road
7	State lands road
8	Other (include toll roads not otherwise indicated or identified above)

SCOUR CRITICAL 113

N	Bridge not over waterway
U	Bridge with "unknown" foundation not evaluated for scour. Risk unknown—monitor during floods and consider closure.
T	Bridge over "tidal" waters not evaluated for scour but considered low risk. Monitor regularly.
9	Foundations on dry land well above flood water elevations
8	Foundations stable; scour above top of footing

7	Countermeasures installed to address previous scour issue; no longer scour critical
6	Evaluation for scour not yet made
5	Foundations stable; scour within footing/pile limits
4	Foundations stable but need protection from further erosion/corrosion
3	Scour critical; foundations unstable for calculated scour conditions
2	Scour critical; extensive scour present, immediate countermeasures required
1	Scour critical; failure imminent. Bridge closed
0	Scour critical. Bridge has failed and is closed

Table 4.2: Categorical bridge features and codes from the National Bridge Inventory Coding Guide [2].

Table 4.3 compares the predictive performance of ordinal encoding and `Word2Vec` embeddings based on accuracy, F1 score, and their respective standard errors (SE). The ordinal encoding model achieved an accuracy of 89.6% (± 1.1%) and an F1 score of 89.2% (± 1.2%), while the `Word2Vec` embedding model slightly outperformed it with an accuracy of 91.5% (± 0.8%) and an F1 score of 91.3% (± 0.7%). The p-values of 0.181 for accuracy and 0.152 for F1 score suggest that the differences between the two encoding methods are not statistically significant, indicating that `Word2Vec` does not provide a clear performance advantage over ordinal encoding in this context.

Metric	**Ordinal Encoding**	`Word2Vec` **Embedding**	***p*-value**
Accuracy ± SE	0.896 ± 0.011	0.915 ± 0.008	0.181
F1 score ± SE	0.892 ± 0.012	0.913 ± 0.007	0.152

Table 4.3: Performance comparison between Ordinal Encoding and Word2Vec Embedding.

Given the nature of our dataset—a small sample size combined with a high number of features—this result is not entirely surprising. High-dimensional feature spaces with limited data can introduce noise, reduce the stability of learned representations, and make it difficult for complex embeddings like `Word2Vec` to offer a clear advantage over simpler encoding methods.

For this reason, we sought to explore a simpler representation to evaluate whether word embeddings could be a valuable direction for further investigation. By experimenting with individual features, we aimed to assess how different encoding strategies performed at a granular level before investigating other embeddings. The results of this feature-level comparison are presented in Table 4.4, where we directly compare the performance of ordinal encoding and `Word2Vec` embeddings across various key features. The first two entries assess the predictive performance of individual features for the target variable *Object Intersected*, while the final entry compares encoding and embedding performance for the feature *Columns per Bent*. These insights will help inform whether word embeddings should be pursued further in future research, either with a larger dataset or additional refinements to embedding techniques.

Target	**Feature**	**Metric**	**Ordinal Encoding**	`Word2Vec` **Embedding**	***p*-value**
Object Intersected	**SERVICE LEVEL 005C**	Accuracy	0.466 ± 0.013	0.511 ± 0.015	**0.036**
		F1 score	0.330 ± 0.013	0.427 ± 0.018	**0.000**
Object Intersected	**ROUTE PREFIX 005B**	Accuracy	0.576 ± 0.012	0.629 ± 0.017	**0.021**
		F1 score	0.515 ± 0.015	0.569 ± 0.018	**0.034**
Columns per Bent	**STRUCTURE TYPE 043B**	RMSE	1.422 ± 0.020	1.360 ± 0.017	**0.030**
		R^2	0.121 ± 0.014	0.188 ± 0.029	0.058

Table 4.4: Comparison of ordinal encoding and `Word2Vec` embedding performance across key features and targets

Table 4.4 presents a comparison of ordinal encoding and `Word2Vec` embeddings across three key features: **SERVICE LEVEL 005C**, **ROUTE PREFIX 005B**, and **STRUCTURE TYPE 043B**. The results highlight how different encoding methods influence predictive performance, with accuracy, F1 score (or RMSE and R^2 for **STRUCTURE TYPE 043B**) and associated standard error (SE) reported for each target variable.

For **SERVICE LEVEL 005C**, the `Word2Vec` embedding notably outperformed ordinal encoding, achieving an accuracy of 51.1% (± 1.5%) compared to 46.6% (± 1.3%) and a highly significant p-value of 0.036. The F1 score similarly improved from 0.330 to 0.427, with a highly significant p-value of 0.000, indicating a robust advantage of the embedding approach for this feature.

For **ROUTE PREFIX 005B**, the `Word2Vec` embedding again outperformed ordinal encoding, with an accuracy of 62.9% (± 1.7%) versus 57.6% (± 1.2%) and a statistically significant p-value of 0.021. F1 score also improved from 0.515 to 0.569, reinforcing the embedding's advantage in capturing meaningful patterns within categorical data.

For **STRUCTURE TYPE 043B**, the evaluation metric used is RMSE, where lower values indicate better predictive performance. The `Word2Vec` embeddings achieved a lower RMSE of 1.360 (± 0.017) compared to 1.422 (± 0.020) for ordinal encoding, with a statistically significant p-value of 0.030, suggesting that embeddings offer a meaningful improvement.

Figure 4.8 shows the t-SNE 2D `Word2Vec` embeddings of **STRUCTURE TYPE 043B** derived from the encodings found in Table 4.5 used to predict *Columns per Bent*. Each point corresponds to a specific bridge design category. The spatial arrangement reveals distinct groupings and outliers, suggesting the model captures meaningful structural patterns. For example, "slab design" and "single or spread box beams or girders" are closely positioned in the upper-right quadrant, implying a shared design context. In contrast, "tee beam" and "segmental box girder" are situated further apart in the lower-right quadrant, reflecting their unique structural properties.

Meanwhile, "thru truss" and "multiple box beams or girders" cluster in the upper-left region, separate from more common girder types like "stringer/multi-beam or girder", which lies closer to the origin. "Girder and floorbeam system" appears in the lower-left quadrant, reinforcing its distinct contextual role. The distribution of points demonstrates that the `Word2Vec` embeddings effectively capture design similarity and specialization across bridge types, providing a robust foundation for downstream predictive modeling.

Code	Description
	STRUCTURE TYPE 043B
1	Slab design
2	Stringer/multi-beam or girder
3	Girder and floorbeam system
4	Tee beam
5	Multiple box beams or girders
6	Single or spread box beams or girders

7	Frame (except frame culverts)
8	Orthotropic
9	Deck truss
10	Thru truss
11	Deck arch
12	Thru arch
13	Suspension
14	Stayed girder
15	Movable lift
16	Movable bascule
17	Movable swing
18	The bridge is a tunnel
19	The bridge is a culvert (includes frame culverts)
20	The bridge uses mixed types
21	Segmental box girder
22	Channel beam
00	Another

Table 4.5: Categorical bridge features and codes from the National Bridge Inventory Coding Guide [2].

4.3.4 Discussion

The results indicate that embedding all features simultaneously did not yield a significant improvement over traditional encoding methods. While word embeddings, such as `Word2Vec`, have the potential to capture semantic relationships between categorical values, the small dataset size and high number of features likely introduced noise, making it difficult for the model to leverage the rich representations provided by embeddings fully. The lack of a clear advantage suggests that we need more data and investigations to confirm whether embedding multiple features simultaneously can deliver statistically and practically significant performance gains.

However, we observed promising improvements in predictive performance for certain categorical variables when embedding individual features. These findings strongly suggest that word embeddings have the potential to be highly beneficial in this context, particularly when applied selectively

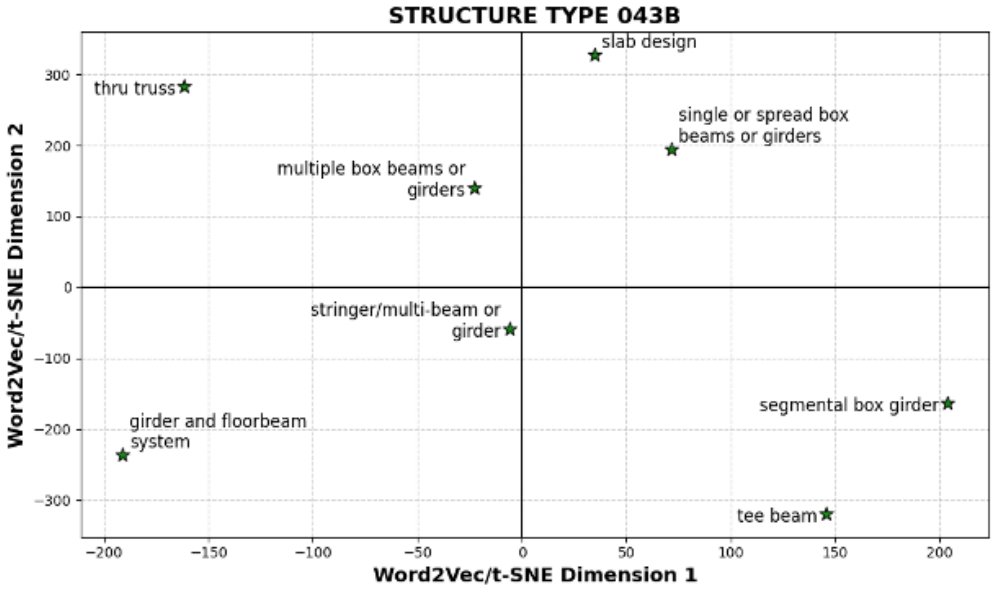


Figure 4.8: Visualization of embeddings for **STRUCTURE TYPE 043B**

to features that exhibit meaningful relationships. Rather than embedding all categorical variables, a targeted approach—focusing on features where embeddings can better differentiate different categories for specific predictive purposes (e.g., different degrees of similarity between varying construction materials matter more for predicting some KBCs than others)—could offer a powerful and interpretable method for encoding categorical data in machine learning models. With more data, this may lead to significant performance gains in predictive modeling for key bridge characteristics.

Beyond the features analyzed in this study, word embeddings could be applied to any of the 50+ categorical variables available in the NBI, especially with more data, refined feature selection, and further validation. Many of these variables encode structural attributes, materials, and functional classifications, which may contain latent relationships that traditional encoding methods fail to capture. By systematically experimenting with embeddings across these additional features, it may be possible to identify patterns and relationships that improve predictive modeling, particularly for complex engineering characteristics that exhibit semantic similarities.

Another opportunity for further refinement is the incorporation of Subject Matter Expert (SME)

feedback into the framework. SME insights could be particularly valuable in evaluating the relative validity of different modeling approaches, offering a critical perspective on method comparisons beyond statistical performance alone. In addition, SME involvement could inform the development of physically meaningful embeddings, guiding the organization of embedded dimensions around interpretable structural characteristics, such as span type, support conditions, or material vulnerabilities. By aligning learned representations more closely with engineering intuition, SME feedback could enhance both the interpretability and the practical applicability of the embedding strategies, ultimately strengthening the bridge between machine-learned models and domain-informed seismic risk assessment.

Furthermore, the NBI dataset provides only short categorical definitions, which may limit the depth of information that can be leveraged for embeddings. However, other external data sources—such as inspection reports, bridge maintenance records, or engineering documentation—may offer richer textual descriptions that could be incorporated into word embeddings. By integrating more detailed text-based datasets, it may be possible to enhance the tabular data available in the NBI, leading to more informative feature representations and ultimately improving model performance. The ability to combine structured and unstructured data through text embeddings presents an opportunity for more comprehensive machine learning models related to bridge seismic vulnerability.

Finally, the landscape of text embedding methods is vast, offering a wide range of techniques that could be explored beyond `Word2Vec`. Approaches such as `FastText`, `BERT`, `GloVe`[86], `XLNet`[112], and OpenAI's transformer-based models [89] provide different advantages depending on the nature of the data and the problem at hand. `FastText`, for example, extends `Word2Vec` by considering subword information, making it more effective for capturing relationships in domain-specific terminology. `BERT` (Bidirectional Encoder Representations from Transformers) and `XLNet` leverage deep contextualized embeddings, meaning they understand word meaning based on surrounding context, rather than relying on fixed word vectors. Transformer-based models such as GPT (from OpenAI) offer embeddings that capture even more complex semantic structures through unsupervised learning on vast text corpora.

One key distinction between embedding techniques is whether they rely on pretrained models or require training from scratch. Pretrained models such as `BERT` and OpenAI's GPT are trained on large-scale corpora (e.g., Wikipedia) and can be fine-tuned for specific tasks, significantly reducing the need for labeled data. These models capture a broad understanding of language and can be directly applied to NBI descriptions or related textual data with minimal adjustment. In contrast, models like `Word2Vec` and `FastText` are often trained on the task-specific dataset itself, making their performance highly dependent on the quantity and diversity of the available text.

In this work, we implemented a pretrained `Word2Vec` embedding model trained on approximately 100 billion words from the Google News corpus, consisting of 300-dimensional word vectors. This allowed us to leverage general-purpose language representations without requiring large volumes of in-domain text. However, because the pretrained model was not trained on structural engineering or transportation-related language, it may miss subtle, domain-specific relationships between terms. In future work, training `Word2Vec` or `FastText` embeddings directly on a large, domain-specific corpus could yield more specialized and informative representations. Exploring these approaches—whether using pretrained models or custom embeddings—offers a promising path for enhancing tabular data in the NBI and improving predictive modeling and structural assessments.

4.4 Uncertainty quantification

Uncertainty quantification (UQ) in machine learning is essential for assessing the confidence in model predictions, ensuring they are both reliable and actionable. By quantifying uncertainty, we can improve model interpretability, prioritize data collection, and support robust decision-making, especially in safety-critical applications. In Chapter 3, we used machine learning to predict key bridge characteristics that influence the calculation of the *Shear-Critical Column Index (SCCI)*. Because these characteristics are predicted rather than directly measured, their uncertainty propagates through the *SCCI* calculation, potentially affecting structural safety assessments. To address this, we introduce a framework for implementing conformal prediction for both classification and regression tasks to quantify uncertainty across a range of bridge parameters. We then apply this framework to key bridge characteristics—including *Columns per Bent*, *Column Minimum Aspect*

Ratio, *Longitudinal Reinforcement Ratio*, *Transverse Reinforcement Ratio*, and *SCCI*—establishing prediction sets and intervals that provide a more comprehensive understanding of potential variability. Conformal prediction enables more informed evaluations of retrofitting needs and safety margins for bridge structures by offering a range of possible outcomes rather than a single point estimate.

4.4.1 Background and method

To guide our approach, Figure 4.9 presents an overview of the analysis process. We begin by identifying the best-performing models from the target variables that we analyzed in Chapter 3. As previously discussed, our AutoML algorithm, AutoGluon, achieved the highest accuracy for several key bridge characteristics. However, a limitation of ensemble-based methods like AutoGluon is their difficulty in directly providing individual uncertainty estimates[1]. Since our goal is to quantify prediction uncertainty, we selected the next-best performing model, Random Forest, to generate predictions along with corresponding uncertainty estimates for this analysis.

To implement uncertainty quantification, we utilized `MAPIE` (Model-Agnostic Prediction Interval Estimator) [39], a `Python` library designed to seamlessly integrate with any `scikit-learn` machine learning model. `MAPIE` applies conformal prediction techniques to quantify uncertainty in both classification and regression tasks, allowing us to estimate prediction intervals and sets without modifying the underlying model. Its flexibility enabled us to incorporate uncertainty estimation into our existing framework with minimal adjustments to our workflow.

Conformal prediction is a statistical technique that provides valid prediction intervals or sets while making minimal assumptions about the underlying model. Unlike traditional uncertainty estimation methods that rely on Bayesian inference or model-specific variance calculations, conformal prediction is a model-agnostic approach that works by leveraging past prediction errors to construct confidence-calibrated intervals. These intervals are designed to maintain a predefined coverage level, ensuring that, for a given confidence level (e.g., 90%), the actual value falls within the estimated range approximately 90% of the time [7, 95]. By applying conformal prediction, we can assess the reliability of our machine learning models and ensure that bridge characteristic

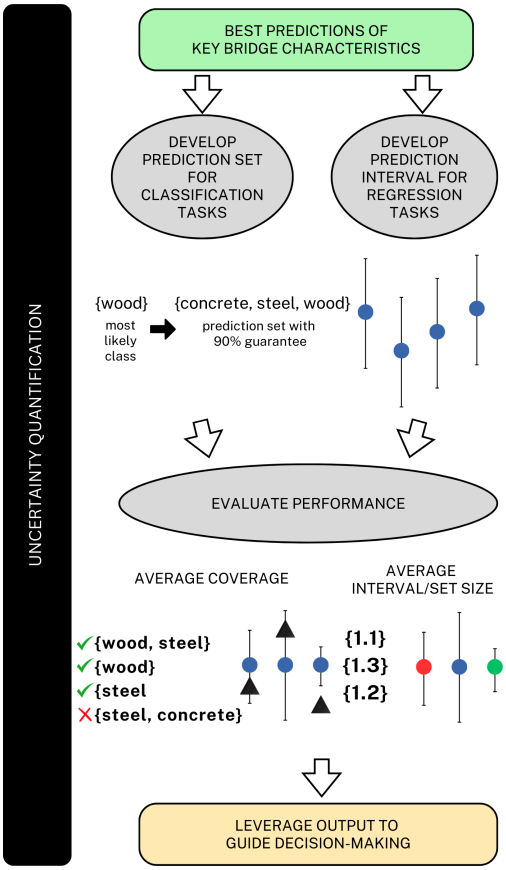


Figure 4.9: Workflow overview for uncertainty quantification using conformal prediction

predictions account for inherent variability and uncertainty in the data-driven models.

In conformal prediction, two key metrics are used to evaluate the reliability of prediction uncertainty estimates: coverage and interval or set size. These metrics provide insight into the trade-off between model confidence and prediction informativeness in both classification and regression tasks. Coverage refers to the proportion of actual values that fall within the predicted set (for classification) or predicted interval (for regression), ensuring that the model's uncertainty estimates are well-calibrated. In classification 90% coverage means that the correct class is included in the prediction set 90% of the time while in regression it means that 90% of actual values fall within the predicted interval. Ideally, coverage should match the chosen confidence level, as excessively high coverage may indicate overly conservative predictions, while lower coverage suggests that the model underestimates uncertainty.

Interval size (for regression) and set size (for classification) refer to the width of the prediction interval and the number of classes included in the prediction set, respectively. In classification, a larger prediction set includes more possible labels, increasing the likelihood that the actual class is captured, but at the cost of reduced specificity. In regression, a larger interval indicates greater uncertainty, meaning the model is less confident in its prediction, whereas a smaller interval suggests higher certainty but may risk excluding the actual value. Ideally, intervals and sets should be as small as possible while maintaining the desired coverage level.

We utilized a 90% confidence level (i.e., 90% target coverage probability) and manually selected the conformal prediction technique—naïve, base, jackknife, or min-max[9, 24]—that provided the desired balance between coverage and interval/set size. Selecting an appropriate conformal method is crucial, as different techniques yield varying levels of conservativeness in their uncertainty estimates. The naïve method, for example, constructs intervals based on residuals from the training data but does not account for variability across different samples, making it more susceptible to under- or overfitting. The base method (split conformal) offers a more stable alternative by using a separate calibration set, ensuring that prediction intervals remain valid while maintaining efficiency. The jackknife and jackknife+ methods provide more robust estimates by leveraging leave-one-out or cross-validation techniques, at the cost of increased computational complexity. Finally, the min-max

method ensures valid prediction intervals even in the presence of outliers by adjusting for extreme cases. By testing these approaches, we identified the method that yielded a desired balance between coverage and interval/set size (ensuring the actual value was captured with the desired probability while keeping intervals or sets as small as possible, preserving prediction informativeness).

4.4.2 Results

Conformal prediction for Key Bridge Characteristics

Classification tasks In Chapter 3, we applied machine learning to classify bridge configurations as either single-column (0) or multiple-column (1) structures. This classification served as a meaningful input, Z, to the *Shear-Critical Column Index (SCCI)*, as defined in Section 3.7.1 and Equation (4.1). For this task, we used a Random Forest Classifier to predict the binary target, and applied conformal prediction using the `MAPIE` framework across ten independent iterations. Table 4.6 presents the aggregated performance metrics, including mean accuracy, F1 score, coverage, and prediction set size. The model achieved a mean accuracy of 0.892 and an F1 score of 0.931, reflecting consistent and balanced classification performance. A mean coverage of 0.911 indicates that the conformal prediction sets reliably captured the actual class, while the average prediction set size of 1.067 suggests that the model typically produced specific predictions, often identifying a single likely class.

$$\textbf{SCCI} = \frac{\left[1.6\left(\frac{\sqrt{f'_c}}{f_y}\right) + 0.5 \cdot \mathrm{TRR}\left(\frac{f_{yt}}{f_y}\right)\right] \cdot \mathrm{L/H_{MIN}}}{\left[1.3Z\left(\frac{\pi}{4}\right)\eta \cdot \mathrm{LRR}\right]} \tag{4.1}$$

Iteration	Accuracy	F1 score	Coverage	Pred Set Size
Mean	**0.892**	**0.931**	**0.911**	**1.067**

Table 4.6: Conformal prediction classification performance across ten iterations using a Random Forest classifier.

While aggregate metrics such as accuracy and F1 score offer valuable insight into overall model

performance, the true strength of conformal prediction lies in its ability to quantify uncertainty at the individual prediction level, as illustrated in Table 4.7. To balance both perspectives, we report mean metrics (e.g., RMSE, R^2, coverage, prediction set size) averaged across all iterations throughout this section. However, we also examine detailed results from a single representative iteration to provide a more intuitive understanding of model behavior. For consistency and fairness, all visualizations and deeper analyses correspond to Iteration 1. Complete metrics for all iterations are available in the supplemental information found in Appendix C.

Table 4.7 presents the conformal prediction results for the binary classification task of predicting the number of columns per bent, shown here for Iteration 1. Each row reports the bridge index, the actual class, the predicted class, and the conformal prediction set. To aid in interpretation, symbols are used to indicate key outcomes: the plus ($^+$) indicate false positives, the negative signs ($^-$) denote false negatives, and the 'not an element of' ($^\notin$) mark the cases where the actual class was not included in the conformal prediction set. These annotations allow for quick identification of classification and conformal prediction errors specific to this iteration.

Regression tasks Similar to our approach for classification tasks, we applied conformal prediction to key bridge characteristics - *Column Minimum Aspect Ratio* (L/H_{MIN}), *Longitudinal Reinforcement Ratio* (LRR), and *Transverse Reinforcement Ratio* (TRR) - which serve as direct inputs to the *Shear-Critical Column Index (SCCI)* calculation, as defined in Equation (4.1). Table 4.8 presents both performance metrics—RMSE and R^2—as well as conformal prediction metrics—coverage and interval size.

Figures 4.10, 4.11, and 4.12 illustrate Iteration 1 of conformal prediction intervals for *Column Minimum Aspect Ratio*, *Longitudinal Reinforcement Ratio*, and *Transverse Reinforcement Ratio*, respectively. Each plot displays 100 samples, ordered by their actual values to help visualize how prediction uncertainty varies across the output range. Black triangles represent the model's predicted values, red dots indicate the actual observed values, and vertical blue lines show the prediction intervals at the 90% confidence level. In this case, the intervals are relatively consistent in width, suggesting uniform model uncertainty across the samples. As expected, most true values

Bridge Index	Actual	Predicted	Prediction Set	Bridge Index	Actual	Predicted	Prediction Set
847	1	1	[1]	**19**$^{+}$	0	1	[0, 1]
243	1	1	[1]	700	1	1	[1]
403	1	1	[1]	**481**$^{+\notin}$	0	1	[1]
219	1	1	[1]	68	1	1	[1]
84	1	1	[1]	476	1	1	[1]
820	1	1	[1]	707	1	1	[1]
566	1	1	[1]	30	0	0	[0]
278	1	1	[1]	93	0	0	[0, 1]
304$^{-\notin}$	1	0	[0]	610	1	1	[1]
151	1	1	[1]	371	1	1	[1]
516	1	1	[1]	676	1	1	[1]
657	1	1	[1]	135	0	0	[0]
427	1	1	[1]	407	1	1	[1]
409	1	1	[1]	224	1	1	[1]
294	1	1	[1]	378	1	1	[1]
301	1	1	[1]	412	1	1	[1]
411	1	1	[1]	509	0	0	[0]
408	1	1	[1]	188	0	0	[0]
81	1	1	[1]	461	1	1	[1]
155	1	1	[1]	159	0	0	[0]
45	0	0	[0]	**396**$^{-\notin}$	1	0	[0]
502	1	1	[1]	331	1	1	[1]
133	0	0	[0, 1]	216	1	1	[1]
291	1	1	[1]	142	0	0	[0]
632	1	1	[1]	**127**$^{-}$	1	0	[0, 1]
280	1	1	[1]	28	1	1	[1]
467	1	1	[1]	514	1	1	[1]
609	1	1	[1]	392	1	1	[1]
72	0	0	[0]	**193**$^{-\notin}$	1	0	[0]
257	1	1	[1]	**168**$^{-\notin}$	1	0	[0]
96	1	1	[1]	829	1	1	[1]
387$^{+\notin}$	0	1	[1]	153	1	1	[1]
221	1	1	[1]	708	0	0	[0, 1]
510	1	1	[0, 1]	26	1	1	[1]
67	1	1	[1]	**125**$^{-\notin}$	1	0	[0]
121	0	0	[0, 1]	314	1	1	[1]
116	1	1	[1]	487	1	1	[1]
17	1	1	[1]	**343**$^{+}$	0	1	[0, 1]
821$^{-\notin}$	1	0	[0]	351	1	1	[1]
136	0	0	[0, 1]	120	0	0	[0]
373	1	1	[1]	465	1	1	[1]
254	1	1	[1]	583	1	1	[1]
134$^{-\notin}$	1	0	[0]	784	1	1	[1]
515^{+}	0	1	[0, 1]	220	1	1	[1]
404	1	1	[1]	825	1	1	[1]
446	0	0	[0, 1]	212	1	1	[0, 1]
166	0	0	[0, 1]	625	1	1	[1]
9	0	0	[0, 1]	659	1	1	[1]

Table 4.7: Conformal prediction sets for *Columns per Bent*. Symbols indicate prediction errors: $^{+}$ = false positive (Actual = 0, Predicted = 1), $^{-}$ = false negative (Actual = 1, Predicted = 0), $^{\notin}$ = conformal prediction set did not include actual value.

Feature	RMSE	R^2	Coverage	Interval Width
Column Minimum Aspect Ratio	**1.877**	**0.391**	**0.92**	**6.489**
Longitudinal Reinforcement Ratio	**0.006881**	**0.181**	**0.93**	**0.025192**
Transverse Reinforcement Ratio	**0.001837**	**0.746**	**0.92**	**0.006927**

Table 4.8: Conformal prediction performance metrics for regression prediction tasks - *Column Minimum Aspect Ratio*, *Longitudinal Reinforcement Ratio*, *Transverse Reinforcement Ratio*

fall within their respective intervals, with a small number of outliers falling outside, consistent with the 90% coverage target.

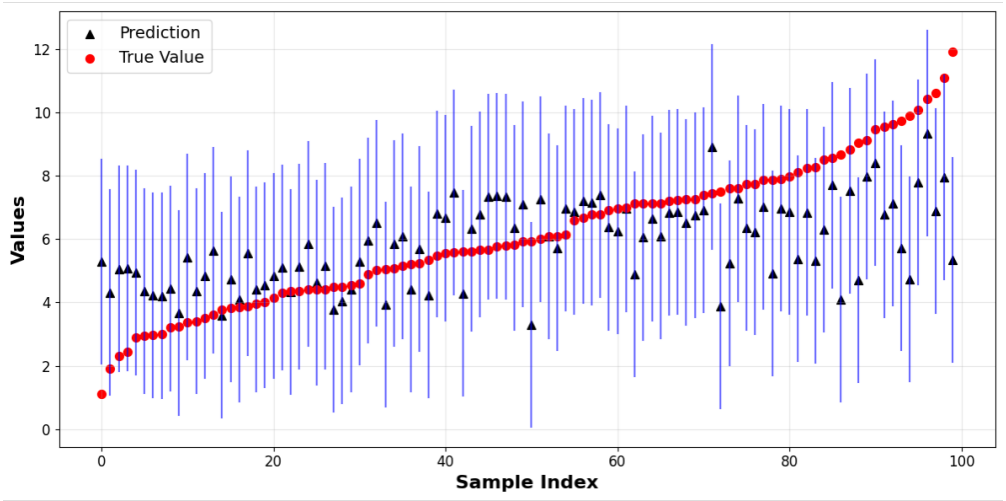


Figure 4.10: Conformal prediction intervals - *Column Minimum Aspect Ratio*

4.4.3 Conformal prediction for Shear-Critical Column Index (SCCI)

Additionally, we analyzed the conformal prediction results for the *Shear-Critical Column Index (SCCI)*. Using the same regression framework, conformal prediction was applied to generate 90% prediction intervals over 10 iterations, using a Random Forest Regressor trained on three different feature sets: Year Built, NBI Only, and NBI + Best KBC Predictions. Table 4.9 summa-

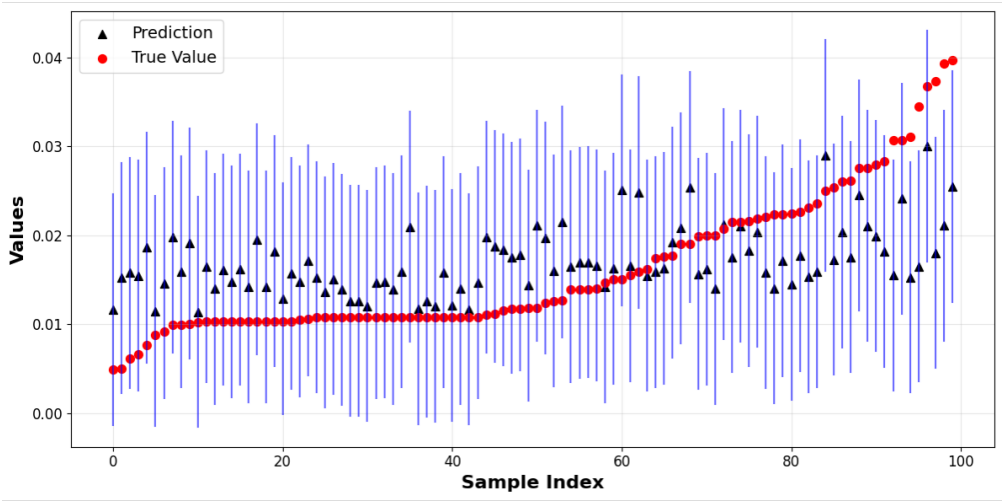


Figure 4.11: Conformal prediction intervals - *Longitudinal Reinforcement Ratio*

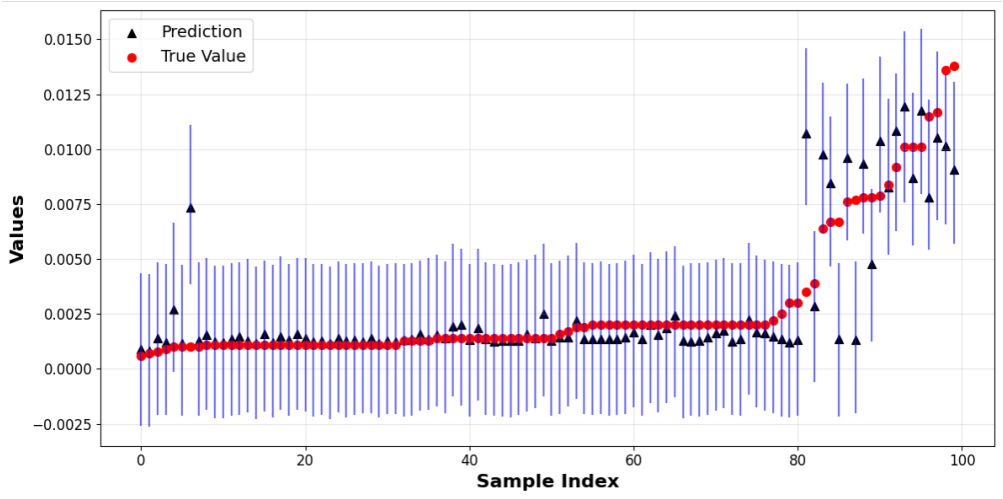


Figure 4.12: Conformal prediction intervals - *Transverse Reinforcement Ratio*

rizes the average performance across these configurations. Compared to the Year Built baseline, both the NBI Only and NBI + Best KBC Predictions models achieved lower RMSE values (0.928 vs. 1.014) and higher R^2 scores (0.174 vs. 0.019), indicating improved predictive accuracy. Coverage remained close to the target level of 90% across all configurations.

These results are consistent with those reported in Chapter 3, where the inclusion of predicted key bridge characteristics alongside NBI data led to modest improvements in model performance. Here, augmenting the NBI features with KBC predictions yields a negligible loss in coverage (0.917 to 0.913; 0.44% reduction) despite a meaningful interval width reduction (i.e., from 3.133 to 3.040; a 2.97% reduction).

Configuration	**RMSE**	**R^2**	**Coverage**	**Interval Width**
Year Built	1.014	0.019	0.919	3.515
NBI Only	0.928	0.174	0.917	3.133
NBI + Best KBC Predictions	0.928	0.174	0.913	**3.040**

Table 4.9: Conformal prediction performance metrics for *SCCI* using three configurations: Year Built only, NBI only, and NBI + Best KBC Predictions.

Figure 4.13 presents the conformal prediction intervals for five randomly selected samples of the *Shear-Critical Column Index (SCCI)* using three different input configurations: Year Built (blue triangles), NBI Only (orange squares), and NBI + Best KBC Predictions (green diamonds). For each sample (indexed along the x-axis), the predicted *SCCI* value is shown as a marker, while the vertical lines represent the corresponding 90% prediction intervals. Red dashed lines indicate the actual *SCCI* values for each sample.

Across most samples, including additional features (NBI and KBC), results in narrower prediction intervals compared to using Year Built alone. This suggests a reduction in predictive uncertainty when more informative input features are included. In several cases, all models produce intervals that contain the actual value, while in others, the prediction intervals vary in their alignment with the actual value *SCCI*. The plot highlights the differences in point predictions and

uncertainty estimates depending on the feature set used.

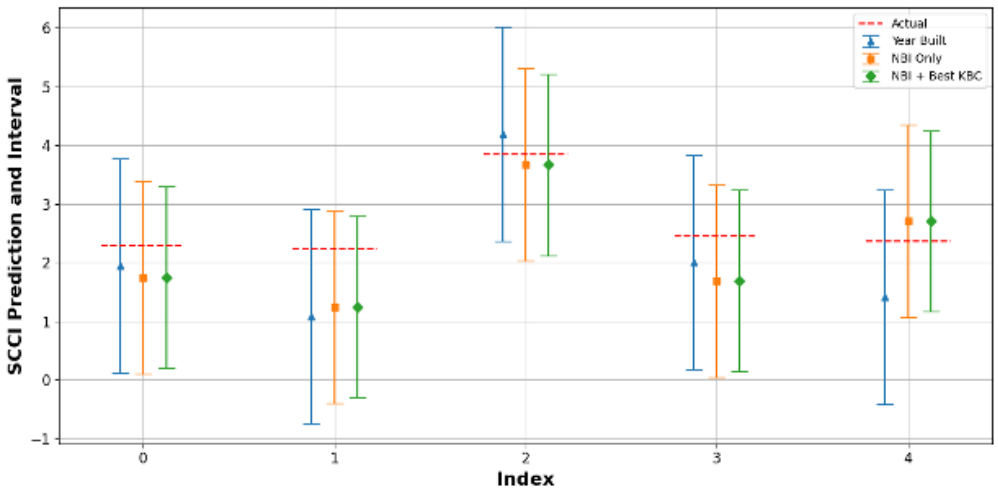


Figure 4.13: Conformal prediction interval comparison for YEAR BUILT, NBI ONLY and NBI + BEST KBC PREDICTIONS feature sets

In Chapter 3, we also aimed to identify bridges with a *Shear-Critical Column Index (SCCI)* below 1, indicating that shear demand exceeds shear capacity. Figure 4.14 presents the conformal prediction results for *SCCI* using theNBI + BEST KBC PREDICTIONS feature set. Actual values are shown as green circles; black triangles represent predicted values; vertical blue lines indicate the 90% prediction intervals. The red dashed line marks the critical threshold of 1.0, which separates shear-critical bridges ($SCCI \leq 1$) from non-shear-critical ones ($SCCI > 1$). The plot shows that many bridges are located near the threshold, making it particularly important to quantify and interpret prediction uncertainty. While most true values fall within their respective prediction intervals, some deviations highlight the practical value of per-sample uncertainty estimates, especially in cases close to the critical decision boundary.

4.4.4 Integrating NLP + UQ: conformal prediction for Object Intersected

In this final analysis, we integrate the Natural Language Processing (NLP) and Uncertainty Quantification (UQ) frameworks to evaluate the reliability of predictions for the **Object Intersected**

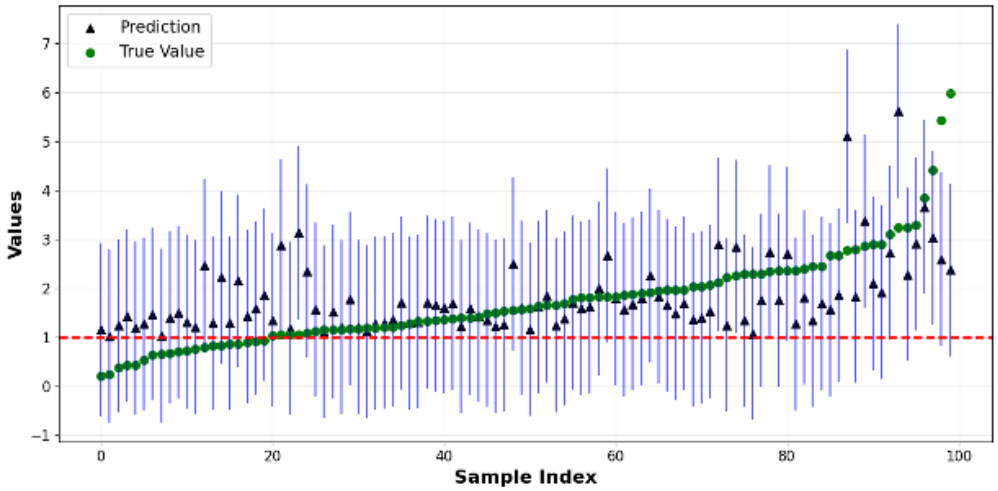


Figure 4.14: Conformal prediction intervals - *Shear-Critical Column Index*

feature. Earlier in the chapter, predictive performance was assessed by comparing ordinal encoding and `Word2Vec`/t-SNE embeddings. Here, we extend the analysis to examine the trustworthiness of the predictions by incorporating conformal prediction to quantify uncertainty.

Model performance is shown in Table 4.10 and was evaluated using both ordinal encoding and `Word2Vec` embeddings. While the differences in overall accuracy were modest, the `Word2Vec` embeddings demonstrated a meaningful advantage in uncertainty quantification. Specifically, the average prediction set size decreased from 1.195 with ordinal encoding to 1.151 with embeddings, indicating more specific predictions under the same coverage threshold (90%).

Sample-level comparisons of the conformal prediction sets (shown in Table 4.11) further illustrate the benefits of embeddings. For example, at Bridge 842, the embedding-based model produced a slightly larger prediction set but successfully included the actual class, providing a more realistic representation of uncertainty compared to the ordinal model, which made an incorrect and overconfident prediction. At Bridges 223, 259, and 236, the embedding model reduced uncertainty by shrinking prediction sets from two classes to one. Conversely, at Bridge 203, the embedding model increased the prediction set size, suggesting a more cautious approach where the ordinal model may have underrepresented uncertainty. Collectively, these results demonstrate that text-based

embeddings not only modestly improve predictive performance but, more importantly, enhance the calibration and interpretability of model uncertainty.

Iteration	Accuracy	F1 score	Coverage	Pred Set Size
Ordinal Encoding	**0.909**	**0.906**	**0.954**	**1.195**
`Word2Vec` **Embedding**	**0.912**	**0.910**	**0.948**	**1.151**

Table 4.10: Conformal prediction classification performance across ten iterations using XGBoost classifier.

Ordinal Encoding				Word2Vec Embedding			
Bridge Index	**Actual**	**Prediction**	**Prediction Set**	**Bridge Index**	**Actual**	**Prediction**	**Prediction Set**
793	Water	Water	[Water]	793	Water	Water	[Water]
493	Underpass	Underpass	[Underpass]	493	Underpass	Underpass	[Underpass]
629	Underpass	Underpass	[Underpass]	629	Underpass	Underpass	[Underpass]
461	Underpass	Underpass	[Underpass]	461	Underpass	Underpass	[Underpass]
602	Underpass	Underpass	[Underpass]	602	Underpass	Underpass	[Underpass]
842	Underpass	Water^{+}	[Water]$^{\notin}$	842	Underpass	Underpass	[Underpass, Water]
739	Water	Water	[Water]	739	Water	Water	[Water]
423	Ramp	Ramp	[Ramp]	423	Ramp	Ramp	[Ramp]
223	Underpass	Underpass	[Ramp, Underpass]	223	Underpass	Underpass	[Underpass]
288	Underpass	Underpass	[Underpass]	288	Underpass	Underpass	[Underpass]
678	Water	Water	[Water]	678	Water	Water	[Water]
663	Water	Water	[Water]	663	Water	Water	[Water]
259	Overpass	Overpass	[Overpass, Underpass]	259	Overpass	Overpass	[Overpass]
447	Underpass	Underpass	[Underpass]	447	Underpass	Underpass	[Underpass]
76	Overpass	Overpass	[Overpass]	76	Overpass	Overpass	[Overpass]
236	Underpass	Underpass	[Ramp, Underpass]	236	Underpass	Underpass	[Underpass]
747	Water	Water	[Water]	747	Water	Water	[Water]
769	Water	Water	[Water]	769	Water	Water	[Water]
25	Underpass	Underpass	[Underpass]	25	Underpass	Underpass	[Underpass]
674	Water	Water	[Water]	674	Water	Water	[Water]
203	Ramp	Ramp	[Ramp]	203	Ramp	Ramp	[Ramp, Underpass]
589	Underpass	Underpass	[Underpass]	589	Underpass	Underpass	[Underpass]

Table 4.11: Comparison of prediction sets for ordinal encoding vs. Word2Vec embedding using cumulated score conformal prediction. $^{+}$ marks incorrect predictions; $^{\notin}$ indicates the prediction set did not contain the true label.

4.4.5 Discussion

Uncertainty quantification improves predictive models' interpretability and practical utility by conveying confidence in their outputs. This capability is particularly valuable for resource allocation

and data validation in the context of regional seismic vulnerability assessments of bridges. As it relates explicitly to the data collection efforts for this work, we initially relied on manual effort to painstakingly extract data from the blueprint. In Chapter 3, we demonstrated how to accelerate that process with machine learning. By incorporating conformal prediction, we can further refine this process by identifying bridges with high uncertainty—those with ambiguous prediction sets or large prediction intervals. Instead of verifying every bridge, researchers can prioritize their efforts where model uncertainty is too large to make a meaningful decision, ensuring that expert validation is directed where it is most needed. This targeted approach optimizes data collection efficiency while improving the dataset's reliability for future analyses.

Building on this integration, this chapter's final analysis demonstrates the advantages of combining NLP-based embeddings with uncertainty quantification to strengthen both interpretability and practical utility. The model becomes more transparent and actionable by capturing meaningful feature relationships through text-derived embeddings and quantifying trust in each output via conformal prediction. This integrated approach enables more intelligent decision-making by allowing targeted expert validation of the most uncertain cases and reducing unnecessary verification for high-confidence predictions. In practice, this facilitates more efficient allocation of inspection resources, enhances trust in automated assessments, and ultimately supports the goal of building safer and more resilient infrastructure systems.

Beyond prioritizing expert effort, conformal prediction also offers a novel diagnostic lens, particularly for classification tasks. In addition to identifying uncertain predictions (e.g., a prediction set of [0,1]), conformal methods reveal when the model is overconfident and wrong, such as predicting [0] with high confidence when the actual class is [1]. These high-confidence misclassifications are especially informative for improving model performance, as they may point to underlying data quality issues, mislabeled samples, or overlooked features. In some cases, these overconfident errors may be even more valuable for debugging and refinement than ambiguous predictions, since they reveal the limits of the model's current understanding. Using conformal prediction as a diagnostic tool, model developers can gain greater insight into when and why failures occur, enabling more targeted model retraining and dataset improvements.

The *Shear-Critical Column Index (SCCI)* plot shown in Figure 4.14 highlights another key advantage of uncertainty quantification: enabling more informed decision-making through predictive intervals rather than single-point predictions. Instead of rigidly classifying bridges as shear-critical or not, we now provide a range of possible values. If a prediction interval falls well above or below the critical threshold of one, fewer resources must be devoted to verification. However, additional scrutiny is necessary to determine structural vulnerability with greater certainty when the range straddles the threshold. This prioritization framework applies to *SCCI* and any key metric used in infrastructure planning, including post-disaster assessments, routine maintenance schedules, and retrofit prioritization. By leveraging predictive intervals, decision-makers can allocate resources strategically, focusing efforts where uncertainty is highest.

More broadly, uncertainty quantification is a powerful tool for optimizing resource allocation across various domains. As models improve with additional data, prediction intervals can be refined to enhance coverage and narrow uncertainty. This ensures that limited time, funding, and expertise are concentrated where they are most needed. For instance, infrastructure inspectors possess specialized knowledge, but their availability is limited. Using uncertainty-aware predictions, we can focus efforts on bridges with the most ambiguous classifications, ensuring expertise is applied to the most challenging cases. Similarly, those contributing to bridge database development (e.g., junior engineers) can leverage these workflows to identify characteristics requiring manual verification. Rather than painstakingly reviewing every detail, they can prioritize uncertain cases, efficiently balancing automation with human expertise.

This work's societal benefits and impacts extend to improved disaster resilience and public safety. The ability to produce more accurate and reliable seismic vulnerability assessments equips local governments, engineers, and emergency planners with crucial insights to prioritize and allocate resources toward the most vulnerable infrastructure. This is especially important in regions with high levels of seismic activity, such as the Pacific Northwest, where bridges serve as key conduits for transportation, emergency response, and economic stability during and after an earthquake. By demonstrating a framework for enhancing the representation and reliability of factors related to seismic vulnerability, this research promotes a proactive and efficient approach to disaster preparedness

and resilience.

Chapter 5

CONCLUSION

As disasters caused by natural hazards continue to grow in frequency and severity, the need for more adaptive, data-driven approaches to preparedness and infrastructure resilience has become increasingly urgent. This dissertation addressed three critical data-centric challenges at the intersection of disaster science and infrastructure risk assessment: the underutilization of unstructured textual data in practical disaster reports, the limited availability of network-level data for evaluating seismic vulnerability of bridges, and the difficulty of efficiently leveraging limited but information-rich datasets. Across these domains, the work presented here demonstrates the potential of data science, particularly natural language processing and machine learning, to enhance the analysis of complex datasets and support the development of data-driven solutions in high-stakes environments.

In Chapter 2, we developed a text mining framework to analyze and extract insights from a large, dense corpus of unstructured practical disaster reports. By synthesizing practitioner perspectives, elevating priorities across strategic, operational, and tactical levels, and emphasizing the role of effective written communication, this approach contributes to more informed decision-making and enhanced disaster preparedness. In Chapter 3, we introduced a machine learning framework for predicting key bridge characteristics relevant to seismic vulnerability by leveraging data from the National Bridge Inventory alongside a limited sample of existing records. This approach offers a scalable solution to infrastructure assessment, significantly reducing dependence on time- and resource-intensive data collection efforts. Finally, we built on these efforts in Chapter 4 by presenting a robust, data-efficient procedure combining advanced feature engineering techniques and uncertainty quantification to improve model reliability and interpretability, and it is broadly applicable to other small, information-rich datasets.

Collectively, the contributions of this dissertation provide three analytically rigorous, adaptable

frameworks designed to improve data-driven decision making in disaster preparedness, infrastructure system assessment, and structural performance evaluation. Each framework integrates a tailored analytical approach, supported by rigorous validation procedures and interpretability tools, to ensure methodological soundness and real-world applicability. We validated these frameworks on complex, real-world datasets within the disaster science domain, showcasing their practical utility and robustness in extracting actionable insights from complex, high-dimensional information. These datasets span diverse sources and represent a wide range of data modalities, underscoring the frameworks' capacity to operate under realistic conditions marked by missing data, noisy inputs, and variable data quality.

Despite the strengths of the proposed frameworks, there are two important limitations to consider. First, a central challenge addressed in this work is the inherent difficulty of developing reliable models with small, high-dimensional datasets. In many machine learning applications, performance gains are typically achieved by increasing the size and diversity of training data. However, additional data are often limited or unavailable in domains like disaster science and infrastructure assessment due to logistical, financial, or temporal constraints. This necessitates a shift toward data-efficient modeling strategies that emphasize advanced feature engineering, careful model selection, and rigorous validation. The frameworks presented here were designed with these constraints in mind, aiming to extract the maximum possible value from limited observations. Second, while interpretability tools were integrated throughout, such as topic modeling, feature importance rankings, and uncertainty quantification, expert input remains essential. The practical interpretation of model outputs, particularly in high-stakes environments, depends on domain-specific knowledge to contextualize findings and translate them into actionable insights. These frameworks provide guidance, but human expertise is critical to ensure their responsible and meaningful application.

Looking ahead, there are significant opportunities to build on this work by incorporating additional data sources, such as additional textual records, remote sensing imagery, and sensor-based monitoring systems, to further enhance the accuracy and adaptability of data-driven models. As the volume and variety of disaster-related data continue to grow, so too does the potential for more sophisticated, multi-modal analyses. The frameworks developed in this dissertation provide a strong

foundation for future research at the intersection of machine learning, infrastructure resilience, and disaster preparedness. In particular, this work paves the way for advancing ML/AI approaches capable of extracting meaningful insights from text-heavy, high-dimensional, and limited-quantity datasets—conditions common in disaster science but often underserved in traditional modeling efforts. By addressing these challenges, this dissertation contributes to current practice and the evolving toolkit of researchers and practitioners working to improve resilience in the face of increasingly complex and uncertain hazards.

BIBLIOGRAPHY

[1] Moloud Abdar, Farhad Pourpanah, Shoaib Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. doi: 10.1016/j.inffus.2021.05.008. URL `https://www.sciencedirect.com/science/article/pii/S1566253521001081`.

[2] Federal Highway Administration. *Recording and Coding Guide for the Structure Inventory and Appraisal of the Nation's Bridges*. U.S. Department of Transportation, Federal Highway Administration, Washington, D.C., 1995. `https://www.fhwa.dot.gov/bridge/mtguide.pdf`.

[3] Kristoffer Albris, Kristian Cedervall Lauta, and Emmanuel Raju. Disaster knowledge gaps: Exploring the interface between science and policy for disaster risk reduction in europe. *International Journal of Disaster Risk Science*, 11:1–12, 2020.

[4] Mohamad Alipour, Devin K Harris, and Laura E Barnes. Pattern recognition in the national bridge inventory for automated screening and the assessment of infrastructure. In *Structures Congress 2017*, pages 279–291, 2017.

[5] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.

[6] Filippos Alogdianakis, Loukas Dimitriou, and Dimos C Charmpis. Data-driven recognition and modelling of deterioration patterns in the us national bridge inventory: A genetic algorithm-artificial neural network framework. *Advances in Engineering Software*, 171:103148, 2022.

[7] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[8] Patricia Anthony, Jennifer Hoi Ki Wong, and Zita Joyce. Identifying emotions in earthquake tweets. *AI & SOCIETY*, pages 1–18, 2024.

[9] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. doi: 10.1214/20-AOS1965. URL `https://doi.org/10.1214/20-AOS1965`.

[10] V. V. Bertero and R. A. Virgolini. Earthquake performance of highway systems. *Journal of the Structural Division, ASCE*, 99(ST5):973–996, 1973.

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[12] Arjen Boin and Paul 't Hart. Organising for effective emergency management: Lessons from research. *Australian Journal of public administration*, 69(4):357–371, 2010.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics*, volume 5, pages 135–146. Association for Computational Linguistics, 2017. doi: 10.1162/tacl_a_00051. URL `https://aclanthology.org/Q17-1010/`.

[14] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[15] Donald E Brown. Text mining the contributors to rail accidents. *IEEE Transactions on Intelligent Transportation Systems*, 17(2):346–355, 2015.

[16] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[17] Halil İbrahim Cebeci and Yasemin Korkut. Determination of the critical success factors in disaster management through the text mining assisted ahp approach. *Sakarya University Journal of Computer and Information Sciences*, 4(1):50–72, 2021.

[18] Anita Chandra, Shaela Moen, and Clarissa Sellers. *What role does the private sector have in supporting disaster recovery, and what challenges does it face in doing so?* Rand Corporation Santa Monica, CA, 2016.

[19] Taeyeon Chang, Seokho Chi, and Seok-Been Im. Understanding user experience and satisfaction with urban infrastructure through text mining of civil complaint data. *Journal of Construction Engineering and Management*, 148(8):04022061, 2022.

[20] Mengdie Chen, Sujith Mangalathu, and Jong-Su Jeon. Bridge fragilities to network fragilities in seismic scenarios: An integrated approach. *Engineering Structures*, 237:112212, 2021.

[21] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. Technical report, LearningSys, December 2015. URL `http://terraswarm.org/pubs/767.html`.

[22] Yejin Choi. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155, 2022.

[23] Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35, 2022.

[24] Thibault Cordier, Vincent Blot, Louis Lacombe, Thomas Morzadec, Arnaud Capitaine, and Nicolas Brunel. Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE library. In *Conformal and Probabilistic Prediction with Applications*, 2023.

[25] Earl E Davis, Tianhaozhe Sun, Martin Heesemann, Keir Becker, and Angela Schlesinger. Long-term offshore borehole fluid-pressure monitoring at the northern cascadia subduction zone and inferences regarding the state of megathrust locking. *Geochemistry, Geophysics, Geosystems*, 24(6):e2023GC010910, 2023.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of

deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[27] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[28] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

[29] Adanma Cecilia Eberendu et al. Unstructured data: an overview of the data of big data. *International Journal of Computer Trends and Technology*, 38(1):46–50, 2016.

[30] Nick Erickson, Jonas Mueller, Shixiang Shane Xi, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020. URL `https://arxiv.org/abs/2003.06505`.

[31] Chao Fan, Ali Mostafavi, Aayush Gupta, and Cheng Zhang. A system analytics framework for detecting infrastructure-related topics in disasters using social sensing. In *Advanced Computing Strategies for Engineering: 25th EG-ICE International Workshop 2018, Lausanne, Switzerland, June 10-13, 2018, Proceedings, Part II 25*, pages 74–91. Springer, 2018.

[32] Federal Highway Administration. History of the interstate highway system, 2006. URL `https://highways.dot.gov/highway-history/interstate-system/50th-anniversary/history-interstate-highway-system`. Accessed: 2025-04-12.

[33] Federal Highway Administration. National Bridge Inventory: Washington State Dataset (2023). `https://www.fhwa.dot.gov/bridge/nbi.cfm`, 2023. Accessed April 2025.

[34] Ronen Feldman and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.

[35] Federal Emergency Management Agency FEMA. Developing and maintaining emergency operations plans: Comprehensive preparedness guide (cpg) 101, version 2.0. U.S. Department

of Homeland Security, 2010. URL https://www.fema.gov/sites/default/files/2020-05/CPG_101_V2_30NOV2010_FINAL_508.pdf. Accessed: 2024-12-02.

[36] Ehsan Fereshtehnejad, Gianluca Gazzola, Pratik Parekh, Chirag Nakrani, and Hooman Parvardeh. Detecting anomalies in national bridge inventory databases using machine learning methods. *Transportation research record*, 2676(6):453–467, 2022.

[37] Graziano Fiorillo and Hani Nassif. Application of machine learning techniques for the analysis of national bridge inventory and bridge element data. *Transportation Research Record*, 2673 (7):99–110, 2019.

[38] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1st edition, 1925. Fisher's Least Significant Difference (LSD) test introduced in Chapter 5.

[39] Théo Fougères, Charles Truong, Pierre Nodet, and Ewen Galic. Mapie: Model-agnostic prediction interval estimator for scikit-learn. *Journal of Machine Learning Research (JMLR)*, 2021. URL https://github.com/scikit-learn-contrib/MAPIE.

[40] Xinyu Fu, Chaosu Li, and Wei Zhai. Using natural language processing to read plans: A study of 78 resilience plans from the 100 resilient cities network. *Journal of the American Planning Association*, pages 1–12, 2022.

[41] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66, 2017.

[42] Gabriela Gongora-Svartzman and Jose E Ramirez-Marquez. Social cohesion: mitigating societal risk in case studies of digital media in hurricanes harvey, irma, and maria. *Risk Analysis*, 42(8):1686–1703, 2022.

[43] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[44] Osman Erman Gungor, Imad L Al-Qadi, and Justan Mann. Detect and charge: Machine

learning based fully data-driven framework for computing overweight vehicle fee for bridges. *Automation in Construction*, 96:200–210, 2018.

[45] May Haggag, Mohamed Ezzeldin, Wael El-Dakhakhni, and Elkafi Hassini. Resilient cities critical infrastructure interdependence: a meta-research. *Sustainable and Resilient Infrastructure*, 7(4):291–312, 2022.

[46] Louis Hickman, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146, 2022.

[47] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. doi: 10.1145/1014052.1014073. URL `http://doi.acm.org/10.1145/1014052.1014073`.

[48] Xi Hu and Rayan H Assaad. Improving the predictive analytics of machine-learning pipelines for bridge infrastructure asset management applications: An upstream data workflow to address data quality issues in the national bridge inventory database. *Journal of Bridge Engineering*, 29(1):04023103, 2024.

[49] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning: Methods, Systems, Challenges*. The Springer Series on Challenges in Machine Learning. Springer, 2019. URL `https://www.automl.org/book/`.

[50] Anand Jeyaraj and Amir Zadeh. Institutional isomorphism in organizational cybersecurity: A text analytics approach. *Journal of Organizational Computing and Electronic Commerce*, 30(4):361–380, 2020.

[51] Jianguo Jia, Wen Liang, and Youzhi Liang. A review of hybrid and ensemble in deep learning for natural language processing, 2024. URL `https://arxiv.org/abs/2312.05589`.

[52] Kim A Johnston, Maureen Taylor, and Barbara Ryan. Emergency management communication: The paradox of the positive in public communication for preparedness. *Public Relations Review*, 46(2):101903, 2020.

[53] Achyuthan Jootoo and David Lattanzi. Bridge type classification: Supervised learning on a modified nbi data set. *Journal of Computing in Civil Engineering*, 31(6):04017063, 2017.

[54] Md. Yasin Kabir and Sanjay Madria. A deep learning approach for tweet classification and rescue scheduling for effective disaster management, 2019. URL `https://arxiv.org/abs/1908.01456`.

[55] Mohammad Zaid Kamil, Mohammed Taleb-Berrouane, Faisal Khan, Paul Amyotte, and Salim Ahmed. Textual data transformations using natural language processing for risk assessment. *Risk Analysis*, 2023.

[56] Mostafa Keikha, Ahmad Khonsari, and Farhad Oroumchian. Rich document representation and classification: An analysis. *Knowledge-Based Systems*, 22(1):67–71, 2009.

[57] Christopher SG Khoo and Sathik Basha Johnkhan. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511, 2018.

[58] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4): 150, 2019.

[59] Guy Lebanon, Yi Mao, and Joshua Dillon. The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research*, 8(10), 2007.

[60] Julia Lensing, Jingwen Wang, Branden Johnson, and Youngjun Choe. Text mining of practical disaster reports: Case study on cascadia earthquake preparedness, 2023. URL `https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-4087`.

[61] Guanyang Liu, Mason Boyd, Mengxi Yu, S Zohra Halim, and Noor Quddus. Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques. *Process Safety and Environmental Protection*, 152:37–46, 2021.

[62] Heng Liu and Yunfeng Zhang. Bridge condition rating data modeling using deep learning algorithm. *Structure and Infrastructure Engineering*, 16(10):1447–1460, 2020.

[63] Yi Liu, Yin Gu, and Hui Zhang. Seismic risk assessment and damage analysis: Emerging trends and new developments. *Journal of Safety Science and Resilience*, 2024.

[64] Huan Luo and Stephanie German Paal. A locally weighted machine learning model for generalized prediction of drift capacity in seismic vulnerability assessments. *Computer-Aided Civil and Infrastructure Engineering*, 34(11):935–950, 2019.

[65] S. Mangalathu, M. Shokrabadi, and H. V. Burton. Aftershock seismic vulnerability and time-dependent risk assessment of bridges. Technical Report 2019/04, Pacific Earthquake Engineering Research Center (PEER), 2019. URL `https://doi.org/10.55461/UUUE2614`.

[66] Md Manik Mia and Sabarethinam Kameshwar. Machine learning approach for predicting bridge components' condition ratings. *Frontiers in Built Environment*, 9:1254269, 2023.

[67] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[68] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013. URL `https://arxiv.org/abs/1301.3781`.

[69] S. Miles. Comparison of jurisdictional seismic resilience planning initiatives. *PLOS Currents Disasters*, 2018.

[70] Saif Mohammad and Peter Turney. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010*

Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 26–34, Los Angeles, CA, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-0204.

[71] Saif M Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier, 2016.

[72] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[73] Hyundong Nam and Taewoo Nam. Exploring strategic directions of pandemic crisis management: A text analysis of world economic forum covid-19 reports. *Sustainability*, 13(8):4123, 2021.

[74] J Nassour, D Leykin, M Elhadad, and O Cohen. Computational text analysis of a scientific resilience management corpus: Environmental insights and implications. *Journal of Environmental Informatics*, 36(1), 2020.

[75] National Academies of Sciences, Engineering, and Medicine. *Evidence-Based Practice for Public Health Emergency Preparedness and Response*. National Academies Press, Washington, DC, 2020.

[76] F. Å. Nielsen. Afinn, 03 2011. URL http://www2.compute.dtu.dk/pubdb/pubs/6010-full.html.

[77] Robert Nisbet, John Elder, and Gary D Miner. *Handbook of statistical analysis and data mining applications*. Academic press, 2009.

[78] NOAA National Centers for Environmental Information (NCEI). U.s. billion-dollar weather and climate disasters (2024), 2024. URL https://www.ncei.noaa.gov/access/billions/. Accessed: 2024-10-31.

[79] Oregon Department of Emergency Management. Cascadia subduction zone. https://www.

oregon.gov/oem/hazardsprep/pages/cascadia-subduction-zone.aspx, 2024. Accessed: 2023-01-19.

[80] Abdelhady Omar and Osama Moselhi. Automated data-driven condition assessment method for concrete bridges. *Automation in Construction*, 167:105706, 2024.

[81] OpenAI. Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt.

[82] Hakan T. Otal, Eric Stern, and M. Abdullah Canbaz. Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 851–859, 2024. doi: 10.1109/CAI59869.2024.00159.

[83] Christine Owen, Ben Brooks, Chris Bearman, and Steven Curnin. Values and complexities in assessing strategic-level emergency management effectiveness. *Journal of Contingencies and Crisis Management*, 24(3):181–190, 2016.

[84] Canberk Ozdemir and Sabine Bergler. A comparative study of different sentiment lexica for sentiment analysis of tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 488–496, 2015.

[85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[86] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

[87] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Elsevier, 1980.

[88] Zunxiang Qiu, Quanlong Liu, Xinchun Li, Jinjia Zhang, and Yueqian Zhang. Construction and analysis of a coal mine accident causation network based on text mining. *Process Safety and Environmental Protection*, 153:320–328, 2021.

[89] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Research*, 2018.

[90] Emery Roe and Paul R Schulman. An interconnectivity framework for analyzing and demarcating real-time operations across critical infrastructures and over time. *Safety Science*, 168: 106308, 2023.

[91] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20, 2010.

[92] Enrico Rubaltelli, Stephan Dickert, David M Markowitz, and Paul Slovic. Political ideology shapes risk and benefit judgments of covid-19 vaccines. *Risk Analysis*, 2023.

[93] Matthias Schonlau and Rosie Yuyan Zou. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29, 2020.

[94] Carl H Schultz, Kristi L Koenig, Mary Whiteside, Rick Murray, and National Standardized All-Hazard Disaster Core Competencies Task Force. Development of national standardized all-hazard disaster core competencies for acute care physicians, nurses, and ems professionals. *Annals of Emergency Medicine*, 59(3):196–208, 2012.

[95] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[96] Michael Siegrist and Heinz Gutscher. Natural hazards and motivation for mitigation behavior: People cannot predict the affect evoked by a severe flood. *Risk Analysis*, 28(3):771–778, 2008.

[97] Julia Silge and David Robinson. *Text mining with R : a tidy approach*. O'Reilly Media, 2017. Safari Books Online.

[98] United States Department of Homeland Security. National response framework, fourth edition. https://www.fema.gov/media-library/assets/documents/117791, 2019. Accessed: 2024-03-11.

[99] U.S. Department of Homeland Security. Homeland Security Advisory Council: Countering Violent Extremism (CVE) Subcommittee Report. https://www.dhs.gov/xlibrary/assets/HSAC_CITF_Report_v2.pdf, 2016. Accessed: 2024-03-11.

[100] U.S. Geological Survey. Cascadia subduction zone marine geohazards, 2020. URL https://www.usgs.gov/centers/pcmsc/science/cascadia-subduction-zone-marine-geohazards. [Online; accessed March 12, 2024].

[101] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[102] Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.

[103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[104] RL Villars. Big data: What it is and why you should care. *IDC (International Data Corporation)*, 2011.

[105] Gisela Wachinger, Ortwin Renn, Chloe Begg, and Christian Kuhlicke. The risk perception paradox—implications for governance and communication of natural hazards. *Risk Analysis*, 33(6):1049–1065, 2013.

[106] Brian D Williams. Understanding where emergency management gets the knowledge to solve the problems they face: Where are we more than 20 years after the ijmed special edition calls

on closing the gap? *International Journal of Mass Emergencies & Disasters*, 39(3):417–433, 2021.

[107] Erik Wood and Sarah K Miller. Cognitive dissonance and disaster risk communication. *Journal of Emergency Management and Disaster Communications*, 2(01):39–56, 2021.

[108] Yazhou Xie, Majid Ebad Sichani, Jamie E Padgett, and Reginald DesRoches. The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake Spectra*, 36(4):1769–1801, 2020.

[109] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[110] Dongyang Yan, Keping Li, Shuang Gu, and Liu Yang. Network-based bag-of-words model for text classification. *IEEE Access*, 8:82641–82652, 2020.

[111] Yingwei Yan, Jingfu Chen, and Zhiyong Wang. Mining public sentiments and perspectives from geotagged social media data for appraising the post-earthquake recovery of tourism destinations. *Applied Geography*, 123:102306, 2020.

[112] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan R. Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5754–5764, 2019. URL `https://arxiv.org/abs/1906.08237`.

[113] Munazza Zaib, Quan Z Sheng, and Wei Emma Zhang. A short survey of pre-trained language models for conversational ai-a new age in nlp. In *Proceedings of the Australasian computer science week multiconference*, pages 1–4, 2020.

[114] Wei Zhai, Wanyang Hu, Zhihang Yuan, and Yantong Li. Examining disaster resilience perception of social media users during the billion-dollar hurricanes. *Natural Hazards*, 120(1): 701–727, 2024.

[115] Zhongheng Zhang, Marcus W Beck, David A Winkler, Bin Huang, Wilbert Sibanda, Hemant Goyal, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine*, 6(11), 2018.

Appendix A

CHAPTER 2 SUPPLEMENTAL INFORMATION

In this manuscript we establish an approach for extracting meaningful insights from a corpus of practical reports. While our focus is on the approach, we were able to uncover many meaningful insights that may be useful to researchers and practitioners in the emergency management field. Below we highlight supporting information and additional insights that were gathered from our corpus.

A.1 Corpus development

The Cascadia megathrust earthquake is expected to make an impact that will exceed any single state's emergency response capacity. As a result, the existing response plans assume extensive regional coordination as well as federal support. In this context, city- and county-level response plans are largely, if not fully, embedded in the state-level plans and thus excluded in our corpus to avoid unduly biasing it (e.g., due to a different number of cities and counties that have published their response plans, which overlap with their respective state-level plan). However, future studies that focus on local or municipal plans may carefully compile the relevant documents for further analysis (e.g., to identify any unique elements that stand out in local/municipal plans compared to the state/regional/federal plan) with a consideration that some local/municipal response plans for earthquakes and tsunamis may be absent or buried in a general (not a hazard-specific) emergency response plan.

A.2 Document and corpus feature analysis

A.2.1 Term frequencies

In [Term frequencies - what are the predominant themes in this corpus?] we describe observations based on the frequency of a term found within the corpus. We note that water and transportation are the only two infrastructure systems that made it to the top 15 while 'electricity' and 'telecommunications' are noticeably absent. This trend's possible explanation is that the corpus frequently mentions 'fuel,' which is the primary energy source during the response phase. The corpus perhaps focuses more on the response than the long-term recovery, thereby placing 'fuel' above the electricity and telecommunication infrastructures.

Additionally, we also note that 'earthquake' and 'seismic' are more frequently discussed than 'tsunami.' But, the frequency of 'tsunami' is not too far behind, indicating the practical reports' attention on the coastal communities subject to tsunami. Similarly emphasized is 'information,' highlighting the practical interest around information (gathering and sharing).

A.2.2 Term relationships

In [Term relationships - what are the predominant themes in this corpus?] we illustrate a network of bi-gram terms used throughout the corpus to show the frequency of their use. In addition to the observations described in the manuscript, we also observe other notable pairs are water and transportation, frequently followed by 'system(s),' highlighting practical interests in the infrastructures as 'system(s),' which comprise multiple parts. In contrast, other infrastructure systems, such as 'natural gas,' 'energy (sector),' 'public health,' and 'mass care,' less frequently co-locate with 'system(s)' or 'infrastructure.'

Other tools may be used for a more in-depth study. For example, the `widyr` package in `R` to calculate correlations among common pairs of words co-appearing within the same document section/chapter (not just in adjacency) and create a similar network graph for interpretation. This type of analysis can reveal whether practical reports tend to consider water and transportation systems together within the same section/chapter potentially because of their dependencies in

vulnerability and resilience (e.g., water mains often co-locate with main roads, requiring coordinated repair efforts post-CSZ earthquake). Similarly, such in-depth studies can help identify whether practical reports are concerned about a specific infrastructure sector (e.g., energy) being operated primarily by the private sector, which may pose a challenge in coordinated recovery efforts post-CSZ [18]. Also, the word correlations may help reveal how practical reports consider (inter-)dependencies between infrastructure systems (e.g., Does the discussion of natural gas system co-appear with other dependent systems in the same section/chapter? Is the discussion of mass care placed in the context of public health infrastructure?). Similarly, the word correlations can help reveal more conceptual hubs, e.g., 'Oregon' (instead of 'Washington' or 'California'), which connects liquid fuel, hospital, transportation systems, and water systems, perhaps due to the Oregon Resilience Plan (ORP)'s more thorough consideration of the systems than other states' similar initiatives [69].

A.2.3 Term frequency - inverse document frequency

In [Term frequency - inverse document frequency of whole corpus - what are each corpus document's unique/characteristic themes?], we use tf-idf to examine differences between three select pairs of documents. Below are the pairs used for analysis.

- Washington state CSZ event exercise AAR:

  `AAR Cascadia Rising WA State_2018.pdf`

  `AAR Cascadia Rising 2022 WA State_2022.pdf`

- State-level resilience planning:

  `Resilient Washington State_2012.pdf`

  `Oregon Resilience Plan Final_2013.pdf`

- State-level transportation systems study through the Regional Resiliency Assessment Program (RRAP):

  `WA State Transportation Resiliency Assessment_2019.pdf`

  `Oregon Transportation Systems Resiliency Assessment_2021.pdf`

In addition to the observations annotated in[Term frequency - inverse document frequency of whole corpus - what are each corpus document's unique/characteristic themes?], we note additional findings when comparing Washington's CSZ Preparedness Exercises from 2016 to 2022. In CR16,

the Department of Social and Health Services formed and led an ad-hoc mass care task force (TF) for *ESF* 6 (note that CR16's operations plan was organized around different TFs, such as JTF-WA and GTF). In CR22, "mass care services" became one of two essential core capabilities, along with "critical transportation." In contrast to CR16, CR22 placed much greater emphasis on mass care service and support for *shelters* and *food/nutrition* for *displaced* survivors and people with Access and Functional Needs (*AFN*).

A.3 Sentiment analysis

In [Sentiment analysis] we use the BING lexicon to examine the sentiment of the corpus and found a predominately negative sentiment throughout.

Figure A.1 demonstrates the use of an alternate lexicon, Finn Årup Nielsen (AFINN) discussed in[Sentiment analysis], for analysis of the corpus. Overall sentiment in this figure is represented by the frequency of the word used throughout the corpus multiplied by the assigned AFINN value. Although not as strong as those shown with the BING lexicon, the overall sentiment still tilts towards the negative end, being led by *catastrophic*, *damage*, and *loss*.

Figure A.2 compares the Bing and NRC lexicons, discussed in [Sentiment analysis] as well as the AFINN lexicon, for a single document–Oregon Resilience Plan 2013. The Finn Arup Nielsen (AFINN) lexicon assigns a score between -5 (negative) and 5 (positive) for each word in the corpus. Further analysis can be done using alternate lexicons that may be more specific to the given dataset[76]. Oregon Resilience Plan is the longest document in the corpus making it most suitable to examine the differences between the lexicons. We removed stop words and then divided the document into 100 word sections (index) for easy comparison across the lexicons. The AFINN and Bing lexicons depict the document as overall negative. In contrast, the NRC lexicon sees the overall positive sentiment except for a few negative spots, which align with strongly negative spots identified in the previous two lexicons. A flipped, overall positive outlook of the NRC lexicon-based sentiment analysis is attributed to A) NRC's greater ratio of positive to negative words ($2312/3324 = 69.6\%$) compared with AFINN ($878/1598 = 54.9\%$) and Bing ($2006/4783 = 41.9\%$) [57] and B) a high agreement of the sentiment assigned to each word between AFINN and Bing

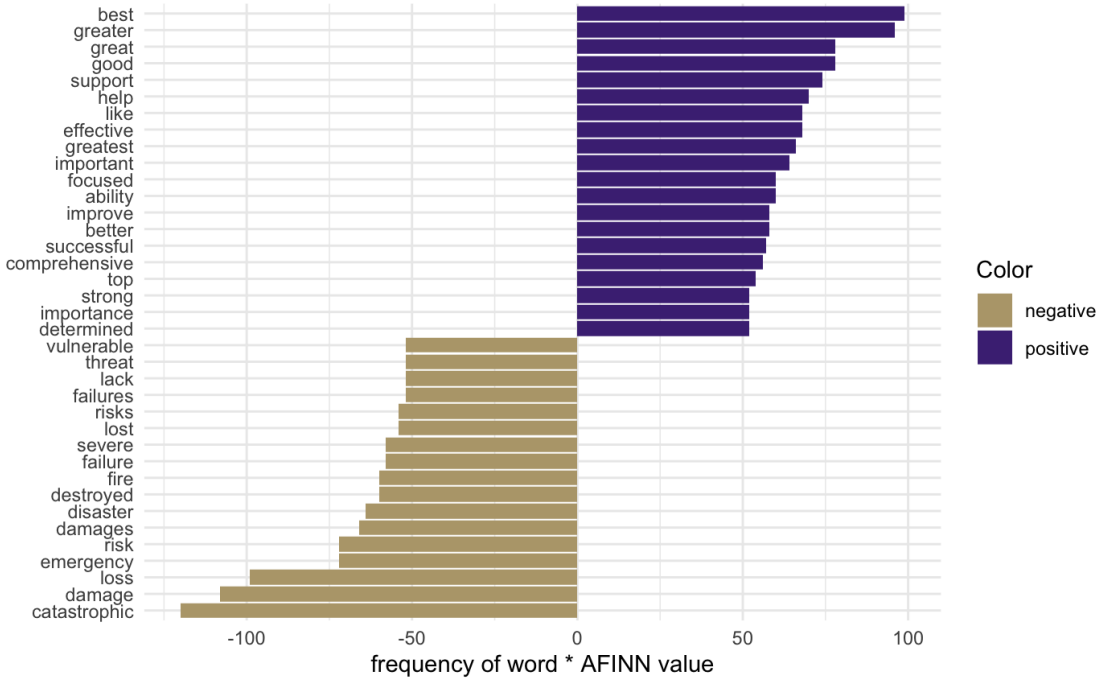


Figure A.1: Sentiment analysis plot using AFINN lexicon

(98.8%) in contrast to NRC agreeing with AFINN at 83.1% and Bing at 80.9% [84].

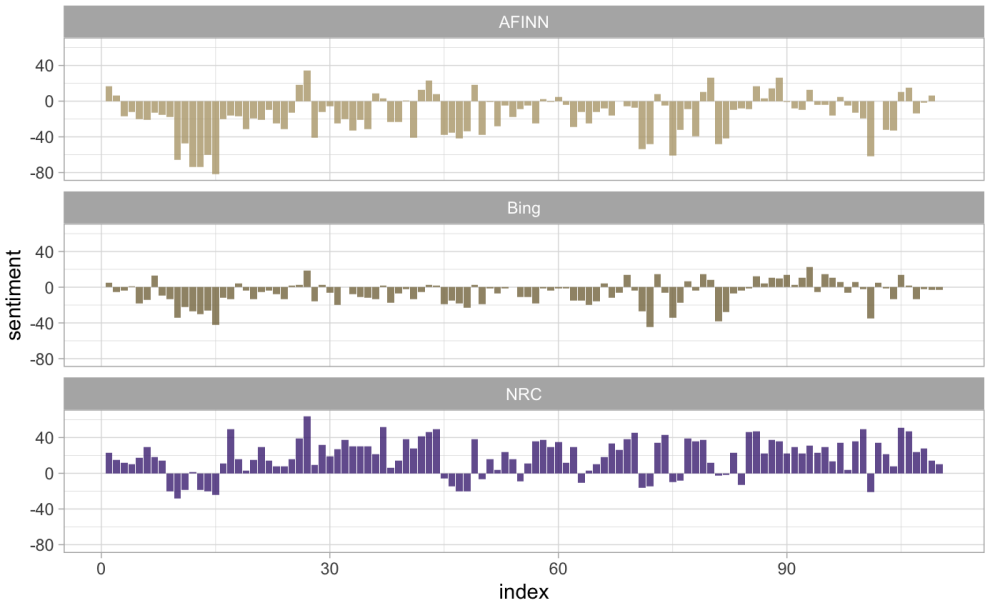


Figure A.2: Comparison of Lexicons based on single document

A.4 Topic modeling

In [Document-topic probabilities using Latent Dirichlet Allocation (LDA) - which documents comprise the identified topics in Figure 2.8?], we examine the use of Latent Dirichlet Allocation (LDA) to conduct topic modeling a 4-Topic model. We also conducted analysis of 2 topics. Figure A.3 for the 2-Topic model suggests that the first dominant topic (on the left) focuses on hazards (e.g., earthquake, seismic, Cascadia [subduction] zone, tsunami) and damages with an emphasis on the State of Oregon and water (infrastructure) system. The second dominant topic (on the right) spans the state-level emergency response and recovery plans that center around fuel, information, support, and exercise. This topic grouping by the model naturally captures two "opposing forces": the first topic group (i.e., hazards and damages) represents what the nature presents to the society whereas the second topic group (i.e., response and recovery plans) represents what is planned by the society in anticipation of the first group. Additionally, there is a notable move of 'recovery'

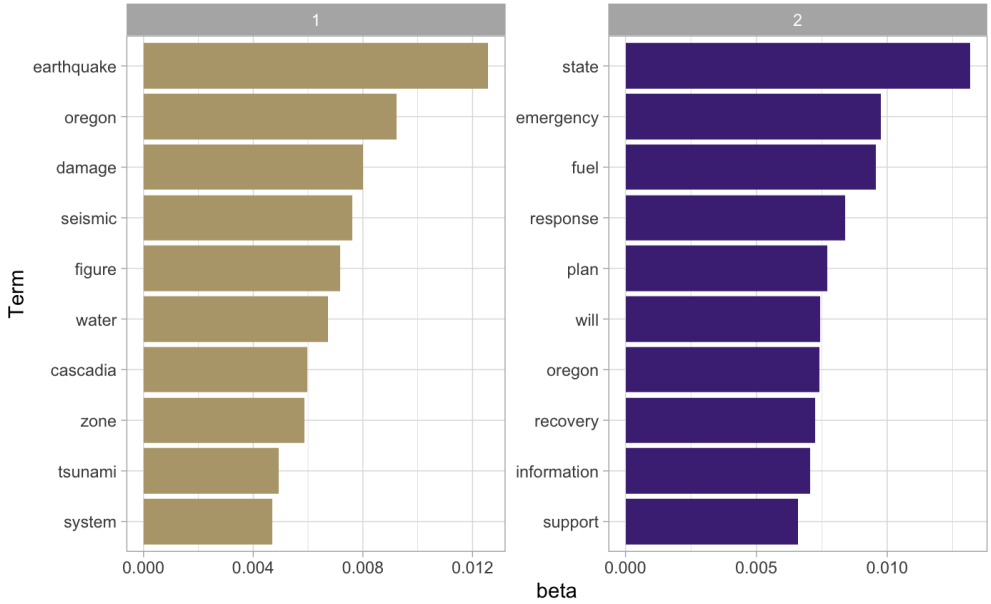


Figure A.3: Practical report topic modeling plot with two topics

from this theme to another theme (i.e., the first group in the 4-Topic modeling) indicates that the concept of 'recovery' appears more frequently with the CSZ EQ, possibly carrying a more general sense than what the concept of 'response' does in the corpus.

Additionally, in [Document-topic probabilities using Latent Dirichlet Allocation (LDA) - which documents comprise the identified topics in Figure 2.8?], we examine how documents are distributed across the 4 topics identified in Figure 2.8. In Figure A.4 we take an alternate visual approach to see which group of documents make up each topic.

In [Topic-word scores using Bidirectional Encoder Representations from Transformers (BERT) - what are the most common topics discussed in this corpus?], we describe the use of HDBSCAN for topic modeling using BERTopic. In addition to topic representations we also exame the clustering hierarchy. A benefit of HDBSCAN is a more interpretable topic choice thanks to its ability to identify a clustering hierarchy, as shown in Figure A.5. It reveals the closeness between Topics 1 and 4, as well as their relative distance from Topics 0, 2, and 3, all measured in terms of the cosine distance between topic embeddings. Note that the closeness of Topics 0, 2, and 3 makes sense as

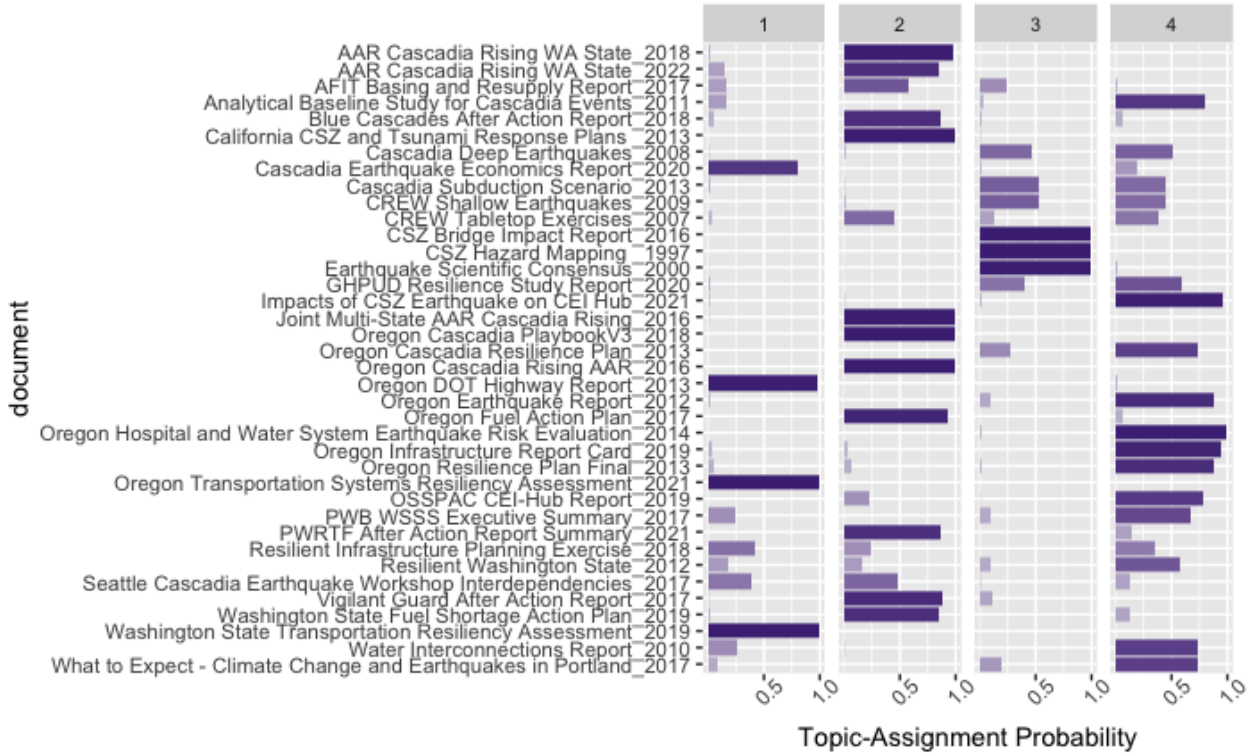


Figure A.4: Topic-assignment probabilities plot using LDA

they generally overlap with Topic 1 (hazards and damages) of the LDA's 2-Topic model, whose Topic 2 (state-level emergency response plan) generally covers BERTopic's Topics 1 and 4. In sum, LDA and BERTopic identified consistent topics within the corpus, demonstrating their robustness, although other corpora may lead to substantially different topics between the two fundamentally different techniques.

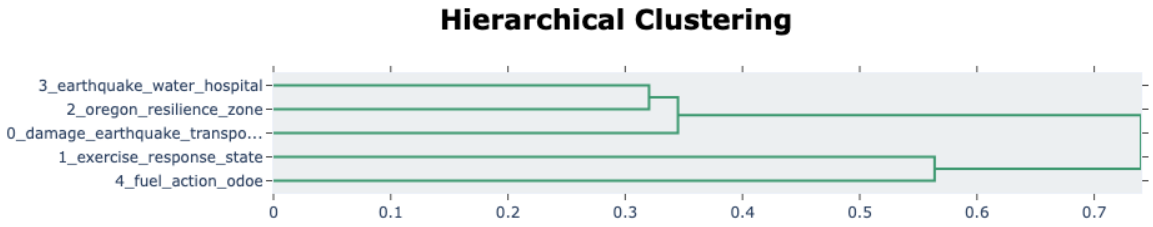


Figure A.5: Hierarchical structure plot by topic

A.5 Text mining of practical disaster reports white paper

Text Mining of Practical Disaster Reports

As you review this set of examples, please consider the following questions.

1. What specific aspects of the tools did you find particularly useful? Why?
2. What challenges or limitations do you see in the tools shown in this document? Why?

What motivated this work?

Throughout the last 20+ years, emergency managers across the Pacific Northwest have conducted planning conferences and emergency response exercises to increase awareness of potential impacts and build preparedness for an M9 Cascadia megathrust earthquake. Emergency management practitioners capture critical lessons learned, actions taken, and planning considerations from these practical field events in after-action reports, response plans, impact assessments, resiliency plans, and other practical reports. These documents are often lengthy and packed with dense information across a wide variety of topics and fields which can make them difficult for a reader to digest. Additionally, these documents are usually geared towards a specific disaster event, locality, or infrastructure which can discourage a reader from reviewing when it does not directly relate to their area of interest. **We propose examining these practical reports as an aggregated collection (corpus) using text mining** in order to increase accessibility and applicability to a wider audience.

What can text mining do?

In this work, we present a suite of text mining tools to efficiently examine this corpus of practical reports. The use of text mining tools can help uncover common trends and themes amongst the vast literature, distinct features that set apart one document from another, emerging patterns throughout a period of time, relationship and connections between documents or represented entities, or underlying emotion across the corpus. Examining this corpus, as an aggregate of information-rich individual reports, may reveal meaningful insights that may not have otherwise been uncovered by single documents alone.

3 categories of text mining tools that we explored in this work.

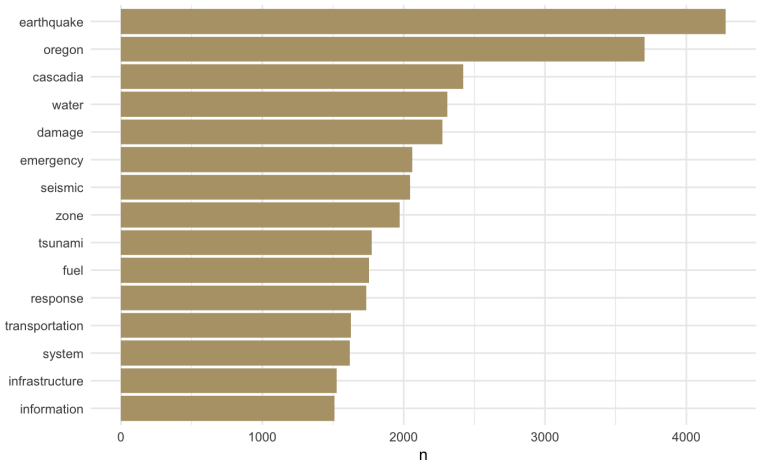
1. Document and Corpus Feature Analysis - examines word frequencies, relationships betweens word, and term frequency-inverse document frequency (tf-idf). These are simple and straightforward tools that aid the user in identifying predominant themes in the corpus.
2. Sentiment Analysis - tool used to identify the sentiment or emotion expressed in the document or corpus based on established lexicons (vocab or set of words specific to a domain or subject).
3. Topic Modeling - assigns a collection of texts into topics/themes using a traditional mathematical model and more recent AI tools like BERTopic and ChatGPT. Topic modeling allows the user to better understand the whole collection, aid in uncovering hidden themes from within the collection, assigning documents to discovered themes, or using these assignments to further organize or summarize the collection.

Document Collection Description

Our corpus consists of 38 publicly available reports representing emergency preparedness planning from Washington State, Oregon, and California as well as regional and federal level planning documents. A list of these reports can be found here. These reports include after-action reports, resiliency assessments, response plans, and planning scenarios related to the Cascadia Subduction Zone and M9 megathrust earthquake. The corpus encompasses both general reports as well as specific reports related to infrastructure recovery (e.g. water, power, transportation, and telecommunications). This corpus does not include reports related to hospital or healthcare emergency preparedness (unless coupled with another lifeline). Additionally, it is limited to technical, practical, and field reports and does not include academic research articles, news articles, blog posts, or non-technical articles (e.g. informational for the general public).

1. Document and Corpus Feature Analysis

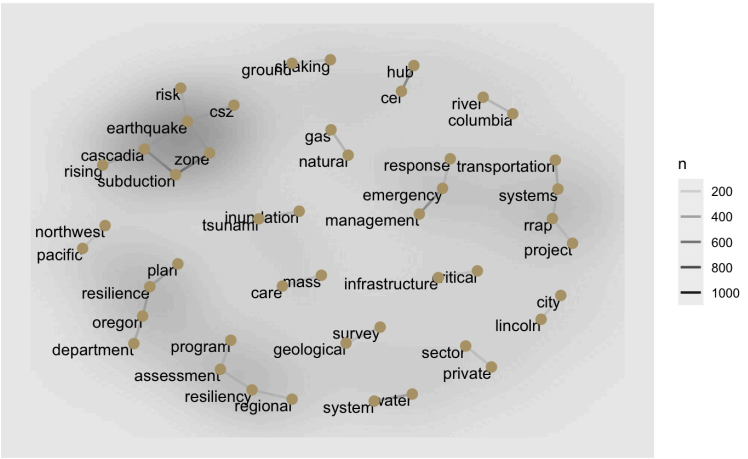
Term Frequency (Figure 1) - Top 15 most frequently used single-word terms in corpus in decreasing order



Probing Question: *what are the predominant themes in this corpus*?

Key Observations: Oregon is the second most frequently mentioned term while Washington did not make it to the list. Only two infrastructure systems (water and transportation) made it to the top 15. Both observations may warrant further investigation into the apparent 'gaps' to determine cause (e.g. if they represent an imbalance in preparedness efforts in practice).

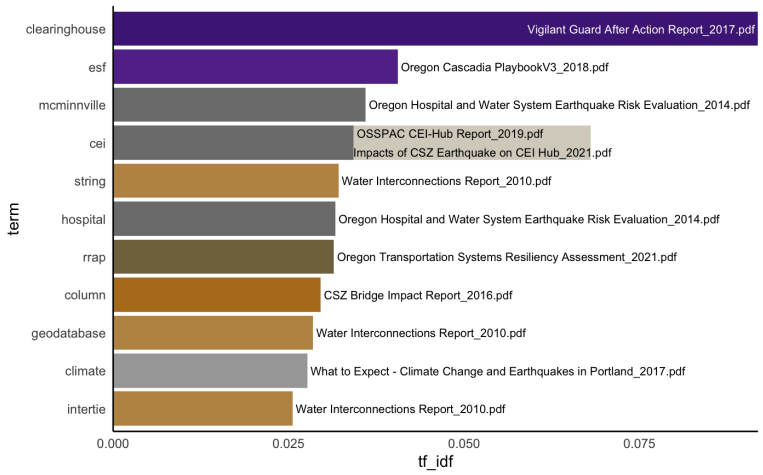
Relationship Frequency (Figure 2) - Relationship between 2 sequential and adjacent terms occurring more than 200 times throughout the corpus. Opacity of the edge and background represent the frequency of the pair.



Probing Question: *what are the predominant themes in this corpus?*

Key observations: Figure 2 reveals more in-depth insights than Figure 1 into how practical reports use each term in its immediate context. (e.g. earthquake, Oregon, and Cascadia, commonly appear in specific contexts).

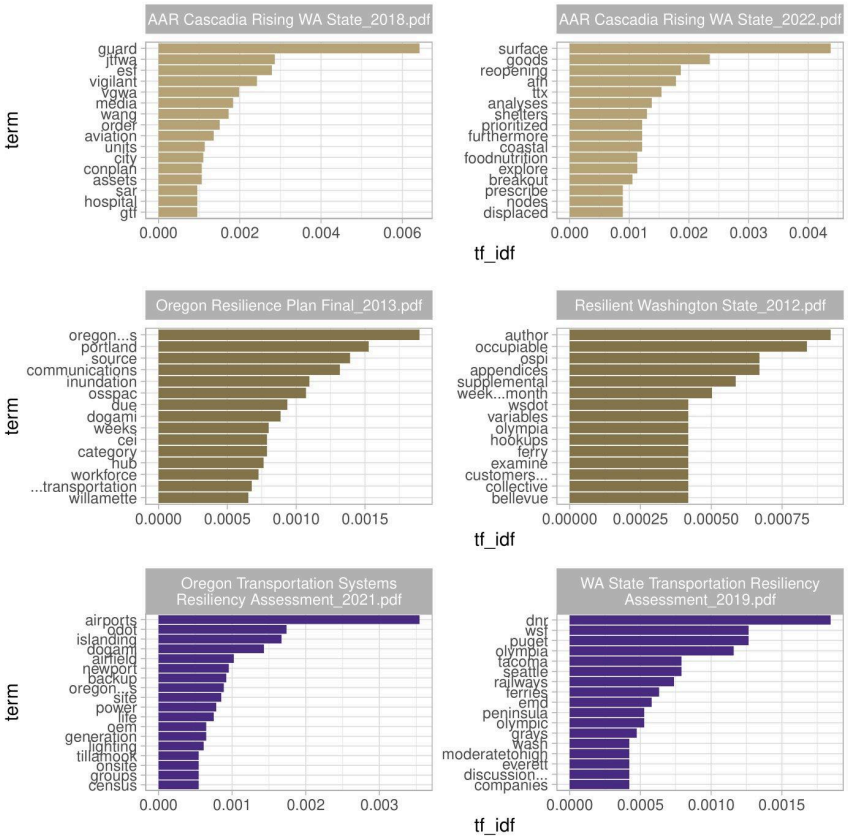
Term Frequency-Inverse Document Frequency (tf-idf) of Whole Corpus (Figure 3) - Top 11 characteristic terms that distinguish each document from the rest in the corpus



Probing Question: *what are the unique/characteristic themes in each document of the corpus?*

Key observations: only one or two documents pay attention to some terms widely considered critical in practice (e.g., clearinghouse, ESF [Emergency Support Function], geodatabase)

Term Frequency-Inverse Document Frequency of Select Pairs (Figure 4) - State level comparison of similar documents. Document pairs: 1) Washington state Cascadia Subduction Zone event exercise AAR in 2018 vs. 2022, 2) state-level resilience planning in Oregon vs. Washington, and 3) state-level transportation systems study in Oregon vs. Washington

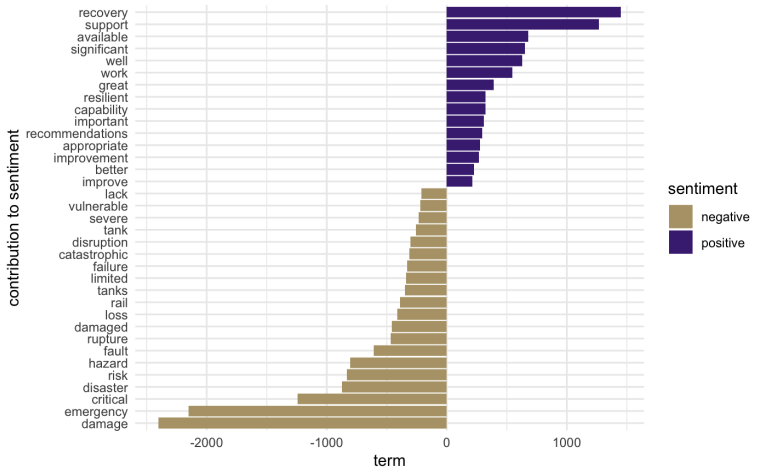


Probing Question: *how do characteristic terms in one document compare to those of another related document?*

Key observations: comparing the tf-idf of these couplets reveals insights into the shift in priorities over time (pair 1) or represented priorities from state to state (pairs 2 and 3).

2. Sentiment Analysis

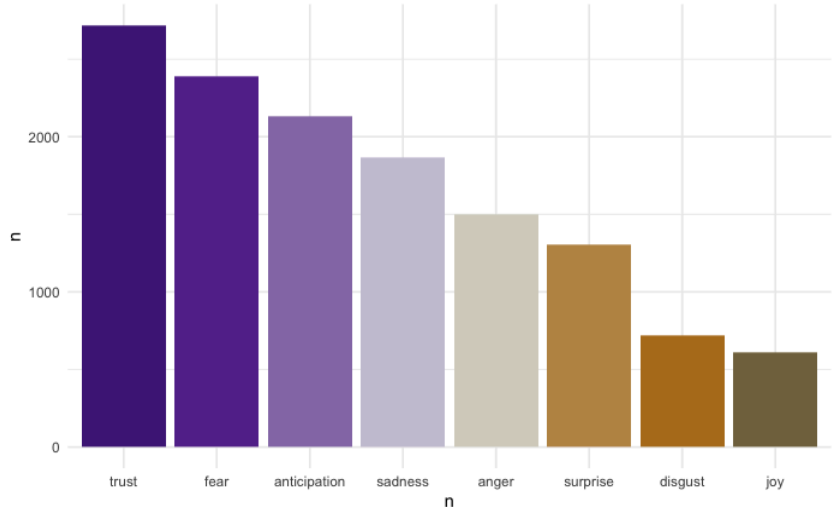
Sentiment Analysis (Figure 5) - term count based on binary (negative/positive) categorization of terms



Probing Question: *what is the portrayed sentiment of this corpus?*

Key Observations: the predominantly negative sentiment of the corpus provides a fresh perspective about the cognitive burden on those who read practical reports in this corpus. This observation may lead authors to reconsider how we present a scenario catastrophic event in practice to induce a more active and constructive engagement of the public in discussion of catastrophe preparedness.

Sentiment Frequency (Figure 6) - mapping of single document to eight basic emotions

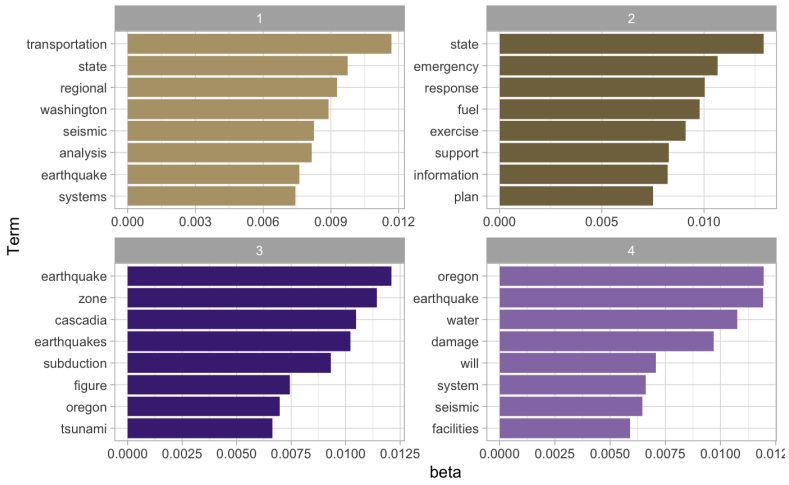


Probing Question: *what emotions are portrayed most in a single document (Oregon Resilience Plan 2013) ?*

Key Observations:
Most frequent are 'trust'-evoking words, followed by 'fear' - evoking words. This analysis shows the potential value of sentiment analysis for better risk communications, where the impacts of trust and negative affect (e.g., fear, sadness, anger) are widely known.

3. Topic Modeling

Topic Modeling with 4 Topics Using Mathematical Modeling (Figure 7) - common topics in the corpus represented by frequently co-occurring terms

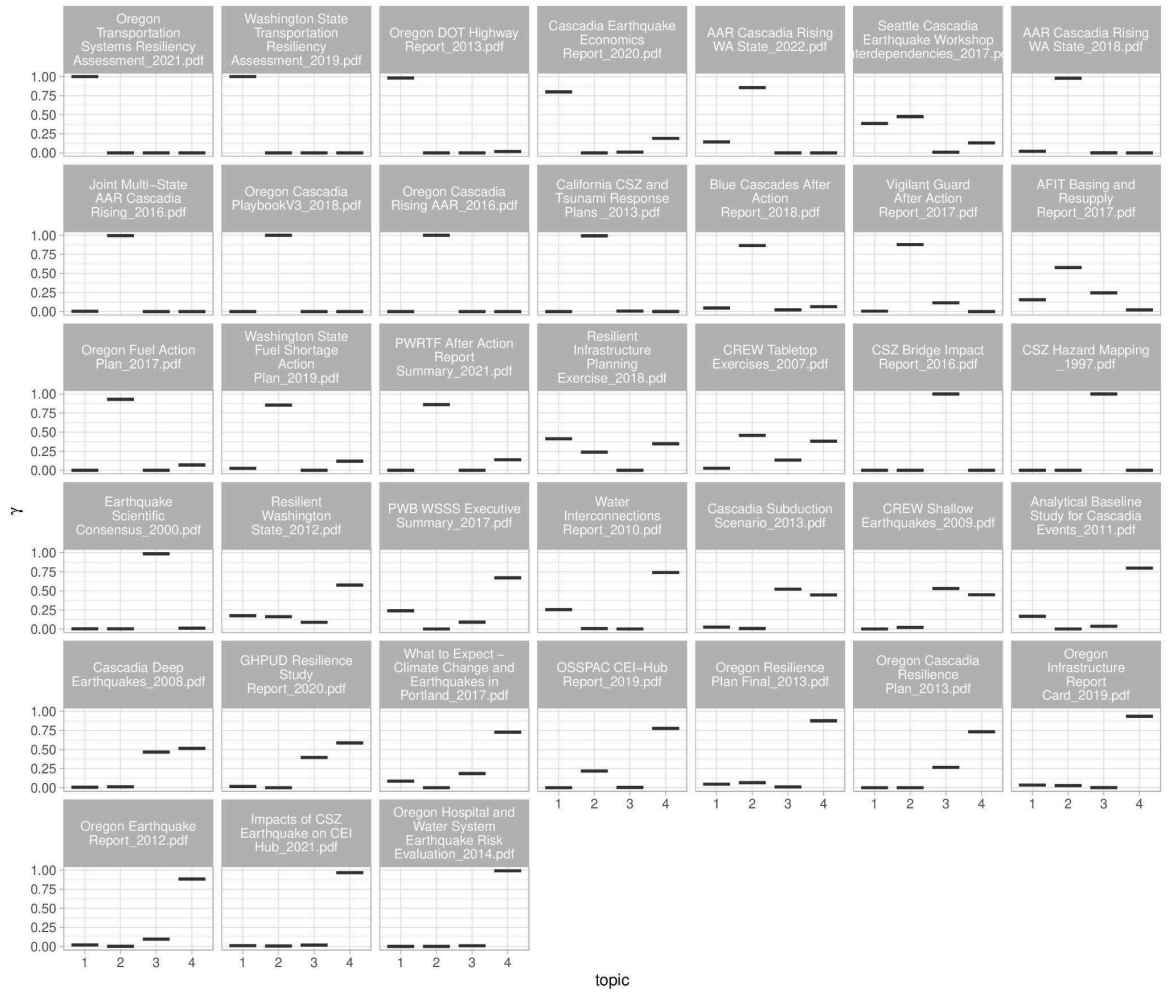


Probing Question: *what are the most common topics discussed in the corpus?*

Key Observations: Topic 1 emphasizes regional and state transportation systems. Topic 2 addresses the state-level emergency response plan around fuel, information, support, and exercise. Topic 3 identifies core hazards and Topic 4 centers around earthquake/seismic damage on water systems/facilities. The ability to specify topics with this method can aid with overall content understanding, information retrieval, and organization of documents.

Note: For this method, the number of topics is specified by the researcher based on domain knowledge and desired outcomes.

Document- Topic Probabilities (Figure 8) - topics (from Figure 7) distributed across the 38 documents in the corpus.

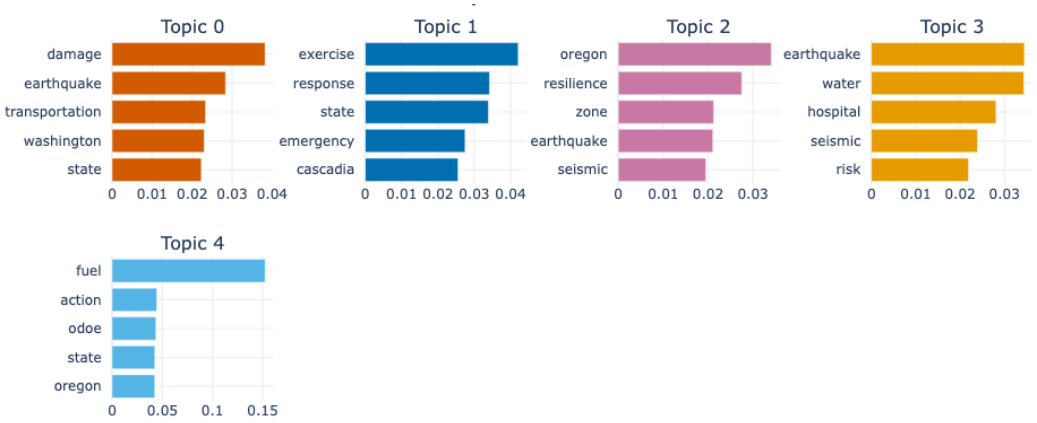


Probing Question: *which documents comprise the identified topics in Figure 7?*

Key Observations: Some documents primarily represent a single topic while others represent a more equal blend of several topics. This type of visualization is generally helpful in summarizing the main themes of the corpus at a high level and in identifying similarly themed documents. Additionally it helps readers to identify relevant documents if exploring a specific topic or theme.

Topic-Word Scores Using BERTopic (Figure 9) - important terms within the identified topic cluster using BERTopic (Bidirectional Encoder Representations from Transformers).

Probing Question: *what are the most common topics discussed in this corpus?*



Key Observations: The resulting five topics (including Topic 0) of BERTopic greatly resemble those found using the mathematical model (Figure 7), but this more advanced AI tool is easier to use because it automatically chooses the number of topics (based on the closeness of terms between and within Topics).

Short Topic Description Using ChatGPT (Table 1) - Brief topic summary using ChatGPT generated from keywords and representative documents from BERTopic Model (Figure 9). Topic (-1) includes outlier documents. Count is the number of representative documents, and representation is the summarization of the representative documents.

Probing Question: *how can you summarize the content of representative documents in each identified topic?*

Topic	Count	Representation
-1	6	Oregon's Infrastructure and the Impact of Cascadia Subduction Zone Earthquake
0	8	Resilient Washington State's Infrastructure Assessment and Recovery Framework
1	7	Cascadia Rising Exercise Response and Planning
2	7	Oregon Resilience Plan for Cascadia Subduction Zone Earthquake and Tsunami
3	6	Earthquake and Seismic Risk Assessment for Oregon's Critical Infrastructure
4	4	Fuel Action Plan in Oregon for Emergency Water Supply

Topic Summaries Using ChatGPT (Table 2) - A longer topic summary (100 words) using ChatGPT

Topic	Count	Representation
-1	6	This topic focuses on the potential impact of a Cascadia Subduction Zone earthquake on various aspects of Oregon. The documents highlight the evaluation of infrastructures and the necessary upgrades to withstand such earthquakes. Additionally, there is information about Cascadia Subduction Zone earthquakes, including a scenario of a magnitude 9.0 earthquake. The topic also includes a summary of a Penrose Conference that discussed the tricentennial anniversary of the Great Cascadia Earthquake. Overall, the topic revolves around understanding the hazards posed by earthquakes in Oregon's Cascadia Subduction Zone and the measures that can be taken to mitigate their effects on infrastructure and public safety.
0	8	This topic focuses on the assessment and analysis of the impacts of earthquakes, tsunamis, and other natural hazards on transportation infrastructure in the Pacific Northwest region, specifically in Washington state and Oregon. The documents mentioned highlight the importance of assessing the resiliency and vulnerability of regional bridge systems and port facilities in the face of potential seismic events, such as the Cascadia Subduction Zone earthquake. The analysis includes considerations of damage, emergency response scenarios, liquefaction risk, and the expected impacts on transportation systems and facilities. The goal is to develop strategies and plans to enhance the resiliency and preparedness of the infrastructure in the region.
1	7	This topic focuses on the planning and response efforts for a potential catastrophic earthquake and tsunami event in the Cascadia Subduction Zone. The documents mentioned include reports on the Cascadia Rising 2016 exercise, which involved multiple states and jurisdictions, including California and Washington. These reports discuss the coordination and management of information, operations, and resources at local, state, and federal levels. The goal is to enhance the capability and readiness of agencies and organizations involved in emergency response and recovery. The importance of public awareness, support, and collaboration in mitigating the impact of such a disaster is also highlighted.
2	7	This topic focuses on the Oregon Resilience Zone and its preparation for a potential earthquake in the Cascadia subduction zone. It discusses plans and strategies to enhance the resilience of buildings, infrastructure, and critical facilities, particularly in coastal areas. The topic also highlights the importance of mitigating fuel and oil-related risks, as well as the need for recovery and recovery models. Documents such as the Oregon Resilience Plan and the CEI Hub Mitigation Strategies provide insights into the state's efforts to reduce risk and improve recovery in the vent of a catastrophic earthquake and tsunami. The topic emphasizes the criticality of preparedness and response measures in mitigating the impact of such a natural disaster.
3	6	This topic focuses on the seismic risk and impact of earthquakes in the state of Oregon, particularly in relation to critical infrastructure such as hospitals and water systems. The documents highlight the economic analysis of a potential Cascadia subduction zone earthquake in the Portland metropolitan region, as well as the risks faced by Oregon's critical energy infrastructure hub. It also includes a study on the earthquake risk evaluation of hospitals and water systems in the state. The topic covers various aspects such as the assessment of damage and risks, energy transmission systems, and the potential impact on the city of McMinnville and the Lincoln and Portland areas.
4	4	This topic focuses on fuel action and emergency response plans related to the state of Oregon. It includes keywords such as fuel, action, emergency, water, supply, agencies, and support. The mentioned documents highlight the importance of preparedness and response in the face of fuel shortages and the need for collaboration among water providers in the Portland metropolitan area. The topic also involves assessing information, planning, and coordinating efforts at the county, federal, and agency levels, with a primary mission to ensure an adequate fuel supply, especially during emergencies. The documents referenced were last revised in October 2017 and April 2019.

Key observations: The generated summaries/representations are overall useful in gaining a quick overview of the topics and distilling a large amount of information into a concise summary. However, the artificial intelligence (AI)-generated wordings (i.e., AI's best attempt to predict a sequence of words) can be misleading to the untrained eye as they were shown to have made up terms (e.g., "Oregon Resilience Zone" in Topic 2 in Table 2) and over-generalized some representative documents (e.g., for Topic 0 in Table 1, some representative documents are not part of the Resilient Washington State Initiative). This misrepresentation potential of AI is an active research topic and at the center of ongoing societal debates.

Share your feedback!

Above we described a suite of text mining tools applied to a corpus of practical reports. The tools presented here are not exhaustive, but demonstrate the opportunities for their use in a variety of different corpora. We invite you to provide your feedback at the link provided below on how you might see the use of these tools in your own work. Thank you in advance for your feedback!

https://forms.gle/ukL3sLLysquWthni6

Appendix B

CHAPTER 3 SUPPLEMENTAL INFORMATION

B.1 Feature selection

Features with low variance contribute little useful information to predictive models, often introducing redundancy and reducing model efficiency. To address this, we removed any numeric feature with a variance below the commonly used threshold of 0.01. Only one feature met this criterion: **TOLL 020**, a nominal variable indicating the toll status of the structure.

Features that are highly correlated can introduce multicollinearity into predictive models, which may distort model interpretations and reduce generalizability. To address this, we conducted a correlation analysis to identify pairs of features with a correlation coefficient greater than 0.9. Within each correlated pair, the feature with the higher variance was retained to preserve the most informative signal, while the other was removed to reduce redundancy. Table B.1 shows the feature, feature type, and description of the 16 features that were removed during feature selection.

B.2 Hyperparameters search ranges

As outlined in Section 3.6.2, we trained models using five widely used machine learning algorithms, applying five-fold cross-validation and random search to explore a range of hyperparameters tailored to each target variable. The specific parameter sets used for tuning are detailed in Tables B.2, B.3, B.4, and B.5, reflecting the unique characteristics and modeling needs of each target.

B.3 Results for binary classification of columns per bent

As outlined in Section 3.7.1, our application of the Shear-Critical Column Index (SCCI) required only determining whether a bridge bent contained single or multiple columns. This allowed us to frame the problem as a binary classification task, enabling more effective prediction and selection of the appropriate value of Z for use in Equation (3.1).

Feature	Feature Type	Description
ADT 029	numeric	Average daily traffic volume for the identified route
APPR WIDTH MT 032	numeric	Normal width, measured in meters, of usable roadway approaching structure
ROADWAY WIDTH MT 051	numeric	Minimum distance, measured in meters, between curbs or rails on the structure roadway
OPR RATING METH 063	nominal	Method used to determine operating rating
DESIGN LOAD 031	nominal	Live load for which the structure was designed
BRIDGE IMP COST 094	numeric	Bridge improvement cost
STRUCTURE LEN MT 049	numeric	Length of the structure, measured in meters
ROADWAY IMP COST 095	numeric	Roadway improvement cost
NAV VERT CLR MT 039	numeric	Minimum vertical clearance imposed at the site, measured in meters
TRAFFIC LANES ON 028A	numeric	Number of lanes being carried on the structure
LAT UND REF 055A	nominal	Reference feature for minimum lateral underclearance on right
MIN VERT CLR 010	numeric	Minimum vertical clearance over the inventory route, measured in meters
WATERWAY EVAL 071	ordinal	Appraisal of the waterway opening with respect to passage of flow through the bridge
POSTING EVAL 070	ordinal	Degree that the operating rating is less than the maximum legal load level; may be used to differentiate between codes
INVENTORY RATING 066	numeric	Capacity rating, referred to as the inventory rating
DECK AREA	numeric	Deck area, measured in meters-squared

Table B.1: Categorical feature descriptions

Model	Hyperparameters
DecisionTreeRegressor	criterion: [squared_error, absolute_error] splitter: [best, random] max_depth: [None, 10, 20, 30] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4]
RandomForestRegressor	n_estimators: [10, 50, 100, 200] criterion: [squared_error, absolute_error] max_depth: [None, 10, 20, 30] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4] bootstrap: [True, False]
SVR	svr__C: [0.01, 0.1, 1, 5, 10] svr__epsilon: [0.001, 0.01, 0.05] svr__gamma: [scale, 0.001, 0.01] svr__kernel: [linear, rbf]
XGBRegressor	n_estimators: [50, 100, 200] max_depth: [3, 5, 7, 9] learning_rate: [0.01, 0.05, 0.1] subsample: [0.6, 0.8, 1.0] colsample_bytree: [0.6, 0.8, 1.0] gamma: [0, 0.1, 0.2, 0.3] reg_alpha: [0, 0.1, 1] reg_lambda: [0, 0.1, 1]
MLPRegressor	mlpregressor__hidden_layer_sizes: [(5,), (10,), (5,5), (10,5), (10,10)] mlpregressor__activation: [relu, tanh] mlpregressor__solver: [adam] mlpregressor__alpha: [0.001, 0.01, 0.1, 0.2] mlpregressor__tol: [1e-4, 1e-3] mlpregressor__learning_rate_init: [0.001, 0.01, 0.02] mlpregressor__early_stopping: [True]

Table B.2: Hyperparameter search ranges used for *Columns per Bent*

The modeling workflow for this target follows the same process as used for all other targets. Hyperparameter ranges are provided in Table B.6, prediction results are summarized in Table B.7, and feature importance is illustrated in Figure B.1.

Model	Hyperparameters
DecisionTreeRegressor	criterion: [absolute_error] splitter: [best] max_depth: [5, 10, 15] min_samples_split: [5, 10, 15] min_samples_leaf: [2, 4, 6]
RandomForestRegressor	n_estimators: [10, 50, 100, 200] criterion: [squared_error, absolute_error] max_depth: [None, 10, 20, 30] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4] bootstrap: [True, False]
SVR	svr__C: [0.01, 0.1, 1, 5, 10] svr__epsilon: [0.001, 0.01, 0.05] svr__gamma: [scale, 0.001, 0.01] svr__kernel: [linear, rbf]
XGBRegressor	n_estimators: [50, 100, 200] max_depth: [3, 5, 7, 9] learning_rate: [0.01, 0.05, 0.1] subsample: [0.6, 0.8, 1.0] colsample_bytree: [0.6, 0.8, 1.0] gamma: [0, 0.1, 0.2, 0.3] reg_alpha: [0, 0.1, 1] reg_lambda: [0, 0.1, 1]
MLPRegressor	mlpregressor__hidden_layer_sizes: [(5,), (10,), (5,5), (10,5)] mlpregressor__activation: [tanh, logistic] mlpregressor__solver: [lbfgs, adam] mlpregressor__alpha: [0.00001, 0.0001, 0.001] mlpregressor__tol: [1e-2, 1e-1] mlpregressor__learning_rate_init: [0.0001, 0.0005, 0.001]

Table B.3: Hyperparameter search ranges used for *Column Minimum Aspect Ratio*

Model	Hyperparameters
DecisionTreeRegressor	criterion: [squared_error] splitter: [best] max_depth: [1, 2, 3, 4] min_samples_split: [3, 5, 10] min_samples_leaf: [6, 8, 12] min_weight_fraction_leaf: [0.005, 0.01, 0.05]
RandomForestRegressor	n_estimators: [10, 50, 100, 200] criterion: [squared_error, absolute_error] max_depth: [None, 10, 20, 30] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4] bootstrap: [True, False]
SVR	svr__C: [1, 10, 50, 100] svr__epsilon: [0.00001, 0.0001, 0.001, 0.01, 0.1] svr__gamma: [scale, 0.01, 0.1, 1] svr__kernel: [rbf]
XGBRegressor	n_estimators: [50, 100, 200] max_depth: [3, 5, 7, 9] learning_rate: [0.01, 0.05, 0.1] subsample: [0.6, 0.8, 1.0] colsample_bytree: [0.6, 0.8, 1.0] gamma: [0, 0.1, 0.2, 0.3] reg_alpha: [0, 0.1, 1] reg_lambda: [0, 0.1, 1]
MLPRegressor	mlpregressor__hidden_layer_sizes: [(5,), (10,), (5,5), (10,5)] mlpregressor__activation: [tanh, relu] mlpregressor__solver: [lbfgs, adam] mlpregressor__alpha: [0.00001, 0.0001, 0.001, 0.01] mlpregressor__tol: [1e-3, 1e-2] mlpregressor__learning_rate_init: [0.0001, 0.0005] mlpregressor__early_stopping: [True]

Table B.4: Hyperparameter search ranges used for *Longitudinal Reinforcement Ratio*

Model	Hyperparameters
DecisionTreeRegressor	criterion: [squared_error, absolute_error] splitter: [best, random] max_depth: [None, 10, 20, 30] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4]
RandomForestRegressor	n_estimators: [10, 50, 100, 200] criterion: [squared_error, absolute_error] max_depth: [None, 10, 20, 30] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4] bootstrap: [True, False]
SVR	svr__C: [0.1, 1, 10, 30, 60, 100] svr__epsilon: [0.001, 0.01, 0.1, 0.5] svr__gamma: [scale, 0.0001, 0.001, 0.01] svr__kernel: [rbf, linear]
XGBRegressor	n_estimators: [50, 100, 200] max_depth: [3, 5, 7, 9] learning_rate: [0.01, 0.05, 0.1] subsample: [0.6, 0.8, 1.0] colsample_bytree: [0.6, 0.8, 1.0] gamma: [0, 0.1, 0.2, 0.3] reg_alpha: [0, 0.1, 1] reg_lambda: [0, 0.1, 1]
MLPRegressor	mlpregressor__hidden_layer_sizes: [(5,), (10,), (5,5)] mlpregressor__activation: [tanh] mlpregressor__solver: [lbfgs] mlpregressor__alpha: uniform(0.001, 0.1) mlpregressor__tol: [1e-5, 1e-4, 1e-3, 1e-2] mlpregressor__max_iter: [1000, 2000, 3000, 4000] mlpregressor__early_stopping: [True]

Table B.5: Hyperparameter search ranges used for *Transverse Reinforcement Ratio*

Model	Hyperparameters

DecisionTreeClassifier	`criterion: [gini, entropy]`
	`splitter: [best, random]`
	`max_depth: [None, 10, 20, 30]`
	`min_samples_split: [2, 5, 10]`
	`min_samples_leaf: [1, 2, 4]`

RandomForestClassifier	`n_estimators: [10, 50, 100, 200, 300, 400]`
	`criterion: [gini, entropy]`
	`max_depth: [None, 10, 20, 30]`
	`min_samples_split: [2, 5, 10]`
	`min_samples_leaf: [1, 2, 4]`
	`max_features: [sqrt, log2, None]`
	`bootstrap: [True, False]`

SVC	`svc__C: uniform(0.1, 10)`
	`svc__kernel: [rbf, poly]`
	`svc__degree: [2, 3, 4, 5]`
	`svc__gamma: [scale]`
	`svc__coef0: uniform(0, 1)`

XGBClassifier	`n_estimators: randint(10, 200)`
	`max_depth: randint(1, 20)`
	`learning_rate: uniform(0.01, 0.3)`
	`subsample: uniform(0.5, 0.5)`
	`colsample_bytree: uniform(0.5, 0.5)`
	`gamma: uniform(0, 0.5)`

MLPClassifier	`mlpclassifier__hidden_layer_sizes: [(50,), (100,), (50, 50)]`
	`mlpclassifier__activation: [tanh, relu]`
	`mlpclassifier__solver: [adam, sgd]`
	`mlpclassifier__alpha: uniform(0.0001, 0.01)`
	`mlpclassifier__max_iter: [1500, 2000, 3000]`

Table B.6: Hyperparameter search ranges used for binary classification of *Columns per Bent* modeling

Model	Accuracy ± SE	F1 Score ± SE
Decision Tree	0.834 ± 0.030	0.835 ± 0.025
Random Forest	0.883 ± 0.020	0.882 ± 0.024
Support Vector Machine	0.854 ± 0.020	0.851 ± 0.024
Gradient Boosting	**0.888 ± 0.024**	0.888 ± 0.024
Neural Network	0.838 ± 0.024	0.832 ± 0.042
AutoML	0.882 ± 0.030	**0.922 ± 0.025**

Table B.7: Binary classification of *Columns per bent* model performance metrics (Average accuracy and F1 score with standard errors over 10 iterations)

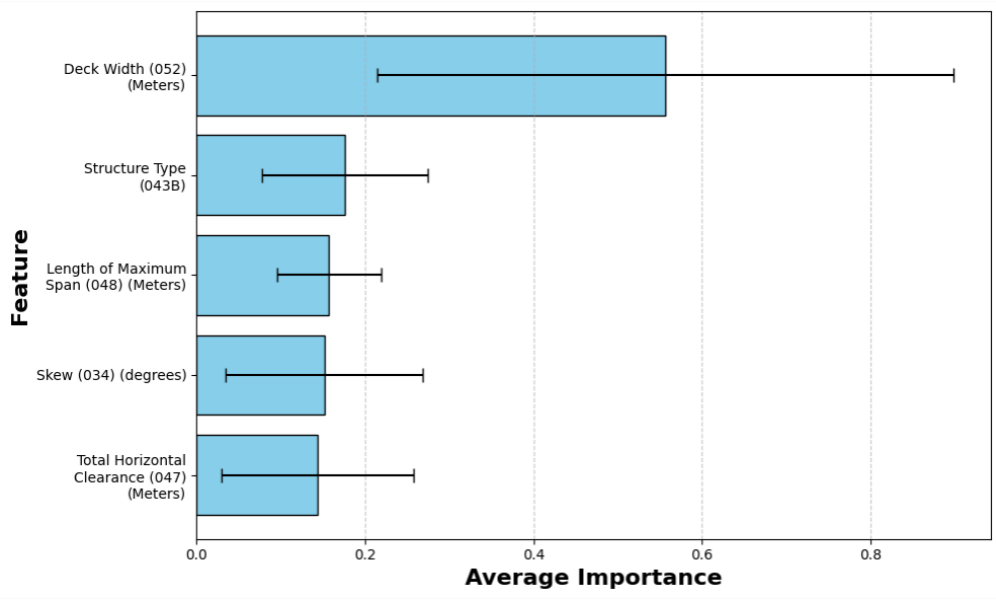


Figure B.1: Permutation feature importance for binary classification of *Columns per Bent*

Appendix C

CHAPTER 4 SUPPLEMENTAL INFORMATION

As described in Section 4.3.2, we also explored a range of general-purpose text embedding methods, including `XLNet`, `SBERT`, `BERT`, and `FastText`. The results of these experiments are presented in Table C.1.

C.1 Additional embedding results

Metric	`XLNet`	`SBERT`	`BERT`	`FastText`
Accuracy ± SE	0.907 ± 0.008	0.901 ± 0.009	0.902 ± 0.011	0.901 ± 0.009
F1 Score ± SE	0.904 ± 0.008	0.895 ± 0.009	0.898 ± 0.011	0.901 ± 0.009

Table C.1: Model performance metrics for predicting *Object Intersected* with embeddings

C.2 Hyperparameter search ranges for Object Intersected

As outlined in Section 4.3.3 and summarized in Table 3.2, we evaluated predictive performance across multiple algorithms using a standardized training and testing framework. Each model underwent hyperparameter optimization through 5-fold cross-validation in conjunction with Randomized Search. The specific parameter ranges explored for each classifier are detailed in Table C.2.

C.3 Feature importance diagram

As discussed in Section 4.3.3, we selected the top eleven features for predicting *Object Intersected* based on a noticeable inflection point in the feature importance scores. As shown in Figure C.1, there is a marked difference in importance beyond the eleventh feature, suggesting that additional features contribute less meaningfully to the model's performance.

Model	**Hyperparameters**

DecisionTreeClassifier	`criterion: [gini, entropy]`
	`splitter: [best, random]`
	`max_depth: [None, 10, 20, 30]`
	`min_samples_split: [2, 5, 10]`
	`min_samples_leaf: [1, 2, 4]`

RandomForestClassifier	`n_estimators: [10, 50, 100, 200, 300, 400]`
	`criterion: [gini, entropy]`
	`max_depth: [None, 10, 20, 30]`
	`min_samples_split: [2, 5, 10]`
	`min_samples_leaf: [1, 2, 4]`
	`max_features: [sqrt, log2, None]`
	`bootstrap: [True, False]`

SVC (with StandardScaler)	`svc__C: uniform(0.1, 10)`
	`svc__kernel: [rbf, poly]`
	`svc__degree: [2, 3, 4, 5]`
	`svc__gamma: [scale]`
	`svc__coef0: uniform(0, 1)`

XGBClassifier	`n_estimators: randint(10, 200)`
	`max_depth: randint(1, 20)`
	`learning_rate: uniform(0.01, 0.3)`
	`subsample: uniform(0.5, 0.5)`
	`colsample_bytree: uniform(0.5, 0.5)`
	`gamma: uniform(0, 0.5)`

MLPClassifier (with StandardScaler)	`mlpclassifier__hidden_layer_sizes: [(50,), (100,), (50, 50)]`
	`mlpclassifier__activation: [tanh, relu]`
	`mlpclassifier__solver: [adam, sgd]`
	`mlpclassifier__alpha: uniform(0.0001, 0.01)`
	`mlpclassifier__max_iter: [1500, 2000, 3000]`

Table C.2: Hyperparameter search ranges for classification of *Object Intersected*

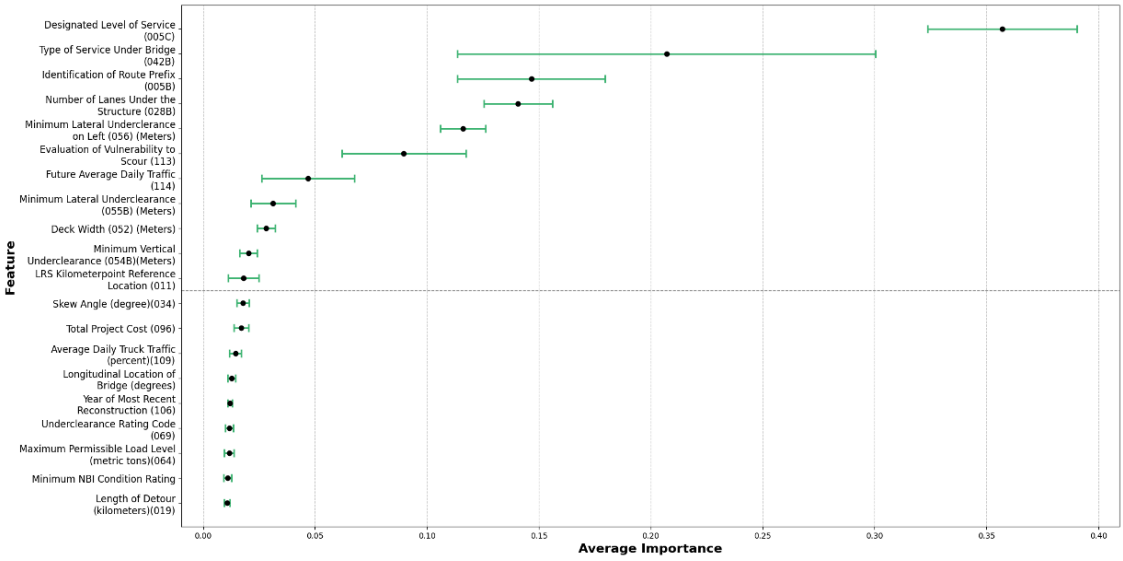


Figure C.1: Feature importance for modeling of *Object Intersected*

C.4 Performance metrics for all iterations

Table C.3 and Table C.4 provide a more detailed breakdown of performance metrics by iteration.

Iteration	Accuracy	F1 Score	Coverage	Pred Set Size
1	0.91	0.941	0.94	1.08
2	0.88	0.923	0.89	1.03
3	0.85	0.903	0.85	1.04
4	0.92	0.949	0.95	1.10
5	0.89	0.931	0.88	1.01
6	0.91	0.943	0.95	1.11
7	0.91	0.944	0.94	1.10
8	0.90	0.936	0.91	1.04
9	0.88	0.923	0.90	1.04
10	0.87	0.913	0.90	1.12
Mean	**0.892**	**0.931**	**0.911**	**1.067**

Table C.3: Conformal prediction classification performance across 10 iterations using a Random Forest classifier.

	Column Minimum Aspect Ratio				Longitudinal Reinforcement Ratio				Transverse Reinforcement Ratio			
	RMSE	**R^2**	**Coverage**	**Interval Width**	**RMSE**	**R^2**	**Coverage**	**Interval Width**	**RMSE**	**R^2**	**Coverage**	**Interval Width**
Iteration 1	1.844	0.411	0.91	6.200	0.007617	0.105	0.88	0.023292	0.001677	0.704	0.94	0.006785
Iteration 2	1.875	0.297	0.92	6.353	0.006325	0.243	0.93	0.024241	0.001637	0.781	0.92	0.007055
Iteration 3	1.971	0.416	0.91	6.805	0.007561	0.111	0.92	0.024330	0.002089	0.758	0.89	0.006878
Iteration 4	1.822	0.398	0.93	6.622	0.007757	0.227	0.88	0.024443	0.002085	0.767	0.90	0.007042
Iteration 5	2.095	0.271	0.89	6.187	0.005979	0.227	0.99	0.026600	0.002145	0.706	0.88	0.006894
Iteration 6	1.936	0.358	0.92	6.553	0.006954	0.287	0.91	0.024916	0.001462	0.815	0.94	0.006957
Iteration 7	1.821	0.441	0.92	6.383	0.005888	0.143	0.98	0.026903	0.001769	0.742	0.93	0.007009
Iteration 8	1.856	0.428	0.93	7.004	0.006918	0.098	0.94	0.026479	0.001934	0.740	0.90	0.006904
Iteration 9	1.669	0.426	0.94	6.428	0.006883	0.124	0.93	0.025457	0.001706	0.729	0.95	0.006948
Iteration 10	1.877	0.465	0.93	6.355	0.006929	0.244	0.94	0.025262	0.001868	0.716	0.92	0.006798
Mean	**1.877**	**0.391**	**0.92**	**6.489**	**0.006881**	**0.181**	**0.93**	**0.025192**	**0.001837**	**0.746**	**0.92**	**0.006927**

Table C.4: Conformal prediction performance metrics for all 10 iterations of regression tasks

Appendix D

PRACTICAL TIPS AND TRICKS FOR APPLYING MACHINE LEARNING TO BRIDGE INFRASTRUCTURE DATA

This appendix provides practical advice and implementation notes gathered throughout the course of this research. The aim is to assist future researchers in reproducing results, troubleshooting common issues, and optimizing workflows when applying the presented frameworks.

Working with the National Bridge Inventory (NBI) and Key Bridge Characteristic Datasets

- **Check for updates to the National Coding Guide.** The NBI evolves over time, and updates to the coding guide may introduce new categories, revise existing definitions, or clarify how missing or ambiguous values should be handled. Always verify that your dataset and variable interpretations align with the most current documentation to avoid misclassification or misinterpretation of bridge attributes.
- **Use ordinal encoding instead of one-hot encoding.** Given the large number of categorical features in the NBI, one-hot encoding can drastically inflate dimensionality, especially for variables with dozens of levels. Ordinal encoding reduces feature count and improves computational efficiency while still capturing category-level information—particularly when combined with models that can handle categorical splits or embeddings.
- **There are many highly correlated features in the NBI.** These redundancies can obscure model interpretation and inflate the perceived importance of related variables. Use correlation heatmaps or variance inflation factors (VIF) to identify and remove or consolidate overlapping variables, especially when building interpretable models.
- **Some continuous features benefit from standardization.** Distributions of numerical variables such as span lengths, clearance heights, or reinforcement ratios can vary widely across bridges. Standardizing these features (e.g., using z-scores) can stabilize model training—especially

for models sensitive to scale, such as support vector machines or neural networks. When in doubt, consult the coding guide to understand each feature's expected range.

- **Review all targets for outliers and identify the specific bridges associated with them.** Outliers may indicate true structural anomalies or data quality issues. Before modeling, perform exploratory analysis to flag extreme values, visualize their distributions, and trace them back to the corresponding bridge records. This not only improves model robustness but also helps surface potential errors or noteworthy cases for further investigation.

Selecting and Applying Modeling and Analysis Tools

- **Be mindful of how you establish your training and testing sets.** The choice of splitting strategy, such as random sampling, stratified sampling, or geographic separation, can significantly affect the model's ability to generalize. Ensure that the split reflects the intended deployment scenario to avoid overestimating performance.
- **Due to the dataset's high dimensionality, be wary of overfitting.** High-dimensional datasets like those in infrastructure assessment contain many redundant or irrelevant variables. Without proper regularization or dimensionality reduction, models can easily memorize the training data. Use cross-validation, keep models as simple as possible, and monitor for signs of overfitting.
- **Feature selection is hard.** Selecting the most relevant variables requires a combination of statistical analysis, domain knowledge, and experimentation. Filter-based methods (e.g., correlation thresholds), wrapper methods (e.g., recursive feature elimination), or model-based importance metrics can all be useful, but no single approach is universally optimal. Iterative refinement is often necessary.
- **Exercise caution with AI-powered tools.** Techniques such as topic modeling or summarization may appear to produce coherent patterns, but they can also generate spurious or non-expert-interpretable outputs. It is especially important to validate these results with domain experts or ground truth, as false signals may not be immediately obvious.

Using AutoML Tools

- **Manual preprocessing often outperforms AutoML defaults.** Although tools like AutoGluon include built-in preprocessing, we found that manually preparing the dataset, such as encoding, standardizing, and handling missing values, often led to better model performance.
- **Be strategic about model quality settings and training time.** AutoGluon allows users to specify both the quality tier (e.g., `medium_quality`, `high_quality`, or `best_quality`) and the maximum training time. While high-quality settings aim to increase accuracy by adding complexity, we found that this often introduced unnecessary overhead, especially when the training time was constrained. Allowing the model to train without a time limit using the `medium_quality` setting often produced results comparable to or better than those from constrained high-quality runs.
- **Understand the default predictor behavior.** Regardless of individual model performance, AutoGluon's default output is typically a weighted ensemble model (`WeightedEnsemble_L2`). If a single model predictor is preferred—for example, for interpretability or speed—you must explicitly select it from the leaderboard results.