THE UNIVERSITY *of* EDINBURGH
School of Engineering

A multimethod approach to learning from text-based construction failure data

Henrietta R. Baker

Supervisor: Simon D. Smith
2nd Supervisor: Gordon Masterton

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Institute of Infrastructure and Environment
School of Engineering

The University of Edinburgh
April 22, 2021

Personal Statement / Declaration of Authorship

I, Henrietta R. Baker, declare that this thesis titled, "A multimethod approach to learning from text-based construction failure data" and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Henrietta R. Baker, April 22, 2021
[Word count: 70593 words]

Abstract

To be sustainable, the construction industry must learn from, and avoid, repetitive failures. At present, there is heavy reliance on learning from case-studies of catastrophic events and a lack of attention to the more frequent, lower consequence and yet repetitive failures. These smaller failures can have huge cumulative impact.

This is important as the construction sector is worth £113 billion per year to the UK economy (6% of UK GDP) and provides over 2.4 million jobs (7% UK jobs). This impressive contribution is undermined by a large number of construction projects which run over time and over cost. This undermining is all the more damaging for those high profile, often publicly funded, infrastructure programmes which attract severe negative publicity when they run overbudget. While other factors contribute to this overspend (for example, inaccurate tender estimates and scope or design change), previous research found that correcting quality mistakes can account for over 20% of a contract's value.

Another failure of the construction industry is its safety performance. In the 2017/18 fiscal year, the fatal injury rate for those working in UK construction was four times the national average at 1.64 per 100,000 workers. Additionally, the Health and Safety Executive in its 2018 Annual Report estimated that safety injuries on site cost £490M to the UK economy. It is therefore both a moral and economic imperative that the industry is learning to avoid repetitive failure.

There is a wealth of information contained within accounts of more frequent, lower consequence incidents and safety observation reports, which should be used. These reports are collected as part of the lifecycle of the project. However, to date, these data have been inaccessible to traditional analysis techniques due to physical accessibility issues and the format of unstructured text data, requiring time consuming manual analysis.

This project harnessed the potential of modern data science methods, including natural language processing (NLP) and machine learning (ML), to produce automated methods and recommendations for analysing these data for the construction industry. A multi-method approach was applied.

First, a qualitative investigation used semi-structured interviews and thematic analysis to explore failure in the construction industry, with particular attention to present 'learning from failure' practice, human factors and biases.

Second, the text-based construction site failure data was analysed using recent data science methods. This analysis relied upon the insights from the first investigation to inform methodological decisions. It was decided to transform the unstructured text data into structured attributes, using machine learning classification methods, for further analysis. Transforming the unstructured text descriptions in this way allows further analysis methods to be performed. Possible further analyses unlocked by this method include risk analysis, graphical analysis, learning, and finer trend analysis.

Finally, qualitative information from the thematic analysis was used to assess usefulness and form recommendations for industrial application of the data analysis methods employed to develop techniques that allow the capture and analysis of data to measure and mitigate the cumulative impact of smaller failures.

Acknowledgements

I feel a very unusual sensation—if it is not indigestion, I think it must be gratitude.

Attributed to Benjamin Disraeli.

While this quote made me smile, it also left me to ponder on how infrequently we step back and acknowledge those to whom we owe the most gratitude, such that gratitude is an unusual emotion - perhaps to be confused with indigestion. I would therefore like to take this opportunity to acknowledge and thank those without whom this thesis would not exist.

Throughout my academic journey, I have had the pleasure of an exceptional academic supervision team, in the form of Dr. Simon Smith and Prof. Gordon Masterton. These two have been teachers of the highest caliber, displaying patience, wisdom and unending support. Thank you Simon and Gordon!

This remarkable academic support continued throughout my 3-month travel scholarship to the University of Colorado with some local additions. I thank Professor Matthew Hallowell and Dr Siddharth Bhandari for their assistance during my stay. Through this connection to Prof. Hallowell, I connected with Dr. Antoine Tixier who also deserves thanks for welcoming me as a collaborator and a co-author for two papers. This change of scenery to Colorado would not be possible without the award of a John Moyes Lessell's Travel Scholarship from Royal Society of Edinburgh (RSE).

This research held at its heart a desire to incorporate context and understanding into the paths it took. This was greatly facilitated by Mr. Bill Hewlett during his time as my industrial supervisor at Costain. Bill was unfailingly honest and shrewd in his guidance of the research, and he was consistently encouraging of my personal development. Thank you Bill for your support.

None of this would be possible without the funding to do so. I am especially thankful to EPSRC (Engineering and Physical Sciences Research Council) and Costain for funding this endeavour. Costain also provided support throughout the years with regular events and a community of PhD researchers. I want to thank Mr. Tim Embley, Ms. Riza Villamor and team for their help and support.

Another team who supported this research are the Edinburgh students who worked on these data for their various dissertations, aiding in labelling data and developing lines of inquiry. Thank you Ms. Milena Velikova, Mr. Edward Campbell, Ms. Amelia Donovan, Mr. Danyal Samad and Mr. Barney Murray.

My family and friends have also played a huge part in getting me to this point. My parents, Suzanne and Richard, and my sisters have been fabulous, offering light relief or a shoulder to lean on when required. I could not be where I am today without you. Thank you so much.

Finally, perhaps the person who had the single biggest job in bringing this thesis into existence (apart from me), was my husband, Ian. Ian, from the moment I asked you to quit your job to move with me to Edinburgh, you have supported me wholeheartedly. Believing in me when I haven't believed in myself and supporting me when I could not support myself. I know I haven't always been easy in the last 4 years and I want you to know that your encouragement has meant the world to me. Thank you.

My sincerest thanks.

Henrietta R. Baker, April 22, 2021

Contents

List of Figures

List of Tables

Glossary

***n*-grams** are combinations of adjacent tokens which are n tokens long. For example, in the previous sentence, “of adjacent tokens” is a 3-gram. 57

accuracy is a performance metric where equal value is assigned to anything correct. This is equivalent to ‘total agreement’ for manual labelling tasks. 105

agreement measures how many labels annotators agree on - ignoring those which both annotators did not assign a label. 130

artificial intelligence (AI) describes any automated process which mimics human-like behaviour. 54

attributes in this research, attributes refer to fundamental features of a work-site, including activity, objects and site environment, which are identifiable before a failure has occurred. 60

bigrams are combinations of tokens which are two tokens long. For example, in the previous sentence, “adjacent tokens” is a bigram. 60

classes are the possible outcomes of a classification process. 102

deep learning methods are machine learning methods which *“allow computers to learn complicated concepts by building them out of simpler ones”* (Goodfellow, Bengio, and Courville, 2016). If represented graphically, these models have many layers, the number of which is referred to as depth. Hence if a model has multiple layers, it is referred to as *deep*. 55

F1 is a performance metric which is a harmonic average of recall and precision. 104

features are the independent variables which are input into a machine learning model. 108

human-centred machine learning “articulates a core set of values and approaches” (Gillies et al., 2016) which deliberately highlight the influence of the human on the machine learning process. 92

hyper-parameters are external inputs to a machine learning algorithm which have to be input by the analyst. 104

labels are the classes assigned to a specific data point. 102

machine learning (ML) describes a wide range of computer programs that implement algorithms and statistical models to carry out tasks (Hastie et al., 2005). These programs rely on patterns and inference from data, rather than explicit instructions, to achieve their aims (Bishop, 2006). 54

Natural Language Processing (NLP) also known as computational linguistics, is a rapidly developing field dealing with the computer analysis of both written and spoken human language. 57

precision is a performance metric which measures whether a class predicted by the ML algorithm is correct or not. 104

recall is a performance metric which measures the proportion of possible correct classes which were predicted. 104

supervised machine learning methods describe methods which require input data with the desired outputs identified manually in order for the algorithms to learn the relationship between these inputs and the desired outputs. 102

tokens for natural language processing generally include words split on white space, but also may include punctuation or numbers. They could instead be sub-word elements. 57

total agreement measures how many of labels and non-labels annotators agree on. 130

vector space are text representations based on the numerical frequency of unique 'tokens' contained within the training vocabulary. These generally form extremely long, sparse vectors. 57

word embeddings are text representation where each word in the vocabulary is represented as a small, dense vector, in a space of shared concepts. 58

Chapter 1

Introduction

Prolegomenon noun : prefatory remarks
pro · · le · · gom · · e · · non | prō – li – gä – mə—nän

Specifically : a formal essay or critical discussion
serving to introduce and interpret an extended work.
Prolegomenon from https://www.merriam-webster.com/dictionary/prolegomenon

At the beginning of each chapter, you will find a short paragraph, such as this, on the title page. These paragraphs contain informal introductions to the content of each chapter, as well as a snippet of my own reflections on that particular phase of the research journey.

1.1 Motivation and Problem Domain

Learning from mistakes is an integral part of human learning. There are innumerable quotes, business textbooks and self-help guides emphasising the importance and potential of learning afforded from failures. A famous quote from Scottish reformist and author Samuel Smiles (1856) stated: "*We learn wisdom from failure much more than from success. We often discover what will do, by finding out what will not do; and probably he who never made a mistake never made a discovery.*" To this end, 'experiential learning' is a thriving area of educational research and practice, demonstrating the eminence of learning from past events to human development. Avoiding repetitive mistakes seems built into the human psyche. Perhaps it is because this type of learning is so ingrained into the human instinct that it is frustrating to discover that industries and organisations fail again and again to learn from their mistakes, and those of their peers. This research investigation intends to alleviate some of these frustrations by developing methodologies to improve organisational learning from failures on UK construction projects.

The construction sector is worth £113 billion per year to the UK economy (6% of UK GDP) and provides over 2.4 million jobs (7% UK jobs) (Office of National Statistics, 2018). This impressive contribution is undermined by a large number of construction projects which run over their tendered time and cost. This undermining is all the more damaging for those high profile, often publicly funded, infrastructure programmes, some of which become notorious for their overspend. A recent report by the Institution of Civil Engineers (ICE, 2019) notes that nine out of ten projects valued over $1bn go over budget worldwide, and that the UK adheres to this trend.

While a range of factors contribute to this overspend (for example, inaccurate tender estimates and scope or design change), Love and Smith (2019) found, in their review paper, that the reported cost of rework is a significant contributor. While reported values average at 3.5% increase to costs, there were several case studies where this increase accounts for over 20% of a contract's value. Rework (the act of repeating a previously completed task) on a construction project can be attributed to incorrect quality of design, materials or workmanship. Incorrect quality is defined as work which contains a defect - either as defined by standards, product specifications or design.

Another area of improvement are the construction industry's safety statistics. As reported by HSE and Health and Safety Executive (2018) for the 2017/18 fiscal year, the current fatal injury rate for those working in UK construction is four times the national average at 1.64 per 100,000 workers. This equates to around 40 people per year losing their lives. An additional 58,000 workers (1.5 times national average) suffered non-fatal injuries while at work. While these figures had been trending downwards for many years, this decline has plateaued in recent years, as seen in Figure 1.1. Not only is there a moral imperative to protect those working on site but reducing safety injuries on site will also reduce the estimated £490M cost to the UK economy (HSE and Health and Safety Executive, 2018), not to mention commercial benefit to individual companies who lose time and money when a staff member is not working, and get fined.

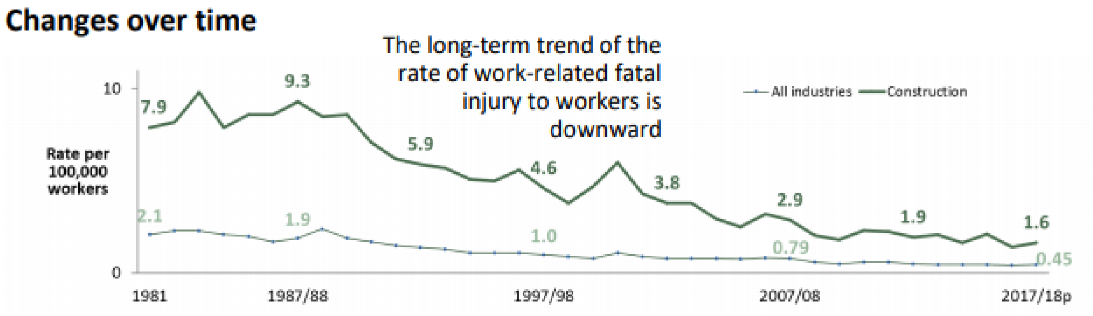


Figure 1.1: Fatality rate over time. Source: NADOR and RIDDOR, 1981 to 2017/18. (HSE and Health and Safety Executive, 2018)

Safety and quality in construction are often analysed separately. This is reflected both in the academic literature and in the industrial organisation. However, there is a strong coupling between the two concepts in the manifestation of a quality failure or safety event. Research undertaken by Wanberg et al. (2013) shows statistically significant positive correlations between safety and quality performance indicators on construction projects. In their research, they spoke to construction employees who postulated that "rework involves demolition, schedule pressure, and unstable work processes" which contributes to more injuries. Also, quality errors (i.e. non-compliance) often cause an unsafe condition, for example, incorrect fixings on a wall panel can result in a falling panel which can injure someone.

The physical occurrences of these failures also have several similarities: they both manifest on site and, along with environmental failures, are immediate. This is elaborated upon later in Section 4.3 in addition to consideration of other construction failure types. For these reasons, both quality and safety failures are considered during this research.

Historically, organisational learning from construction failure has been limited to case-studies of high profile catastrophic failure cases, such as the Hyatt Hotel, Tacoma Narrows Bridge or Charles De Gaul airport walkway collapse. More recent examples include the collapse during construction of a bridge at Florida International University in 2018 resulting in 8 fatalities, the Edinburgh schools construction scandal whose poor quality resulted in the closure of 17 schools for many months, and the Grenfell tower fire which caused 72 deaths. Cases such as these serve as poignant reminders of the risks involved in civil engineering projects, and their detailed forensic analysis has provided vital insights into engineering phenomena and procedures (Breysse, 2012; Delatte, 2010; Pfatteicher and Ph, 2000).

However, as these events are fortunately rare compared to the scale of construction, the learning process is infrequent. Insights gained from the failure analysis of catastrophic case-studies appear most suited towards implementing large legislative changes and procedural upheavals, and not implementing small iterative improvements.

Additionally, the construction industry shouldn't want to wait for a catastrophic event to identify what is going wrong. While transformational change following catastrophic events can invigorate a community, the attention given is often dependent on the severity of the consequence not 'what could have been'. This means that 'warning events' and lower consequence failures may not instigate the change they deserve - even if the cumulative effects are larger than an individual catastrophic event. For a few notorious cases, there are innumerable low profile and legally closed cases from which industry-wide learning is minimal. A method which could anonymise these events for industry-wide trend analysis would help direct resources.

Therefore, we also need a method to implement small iterative improvements. The aviation industry is often praised for their continuous learning processes. Helmreich (2000) states that between 1968–1977 and 2008–2017 the aviation industry achieved a 96% reduction in worldwide death risk per boarding. Like construction disasters, aviation accidents are infrequent yet have potential high loss of life. The same study by Helmreich (2000) noted that an integral part of the aviation learning process is use of data collected about small infractions and observations. Applying this to the construction industry, iterative improvements could be achieved by implementing learning from more frequent, small consequence and yet repetitive failures. Opportunities include using data on NCRs (non-conformance reports) to reduce cost growth through rework and data on small safety incidents to reduce injury occurrences.

On site during the construction phase, safety and quality failure events are, if captured at all, recorded as incident reports, non-compliance reports (NCRs) or near miss events. Before the availability of computers, these reports were physical hard copy forms which would be used reactively on site to implement corrective actions or to file away for posterity. This meant that the data were inaccessible: geographically separate, hard to search and only one copy. However, they are now entered electronically onto computer databases. This makes the data much more accessible, both in term of geography and analysis. Now is the time to think about what digital and data science processes we can implement to help find meaning in these data.

This increase in digital information is not unique to the construction industry. An analysis by the International Data Group (Reinsel, Gantz, and Rydning, 2018) predicts that the amount of digital data will grow from 33 billion Terrabytes (TB) in 2018 to 175 billion TB by 2025. Additionally, much of these data are in unstructured format including audio, video and free-text. This proportion is widely quoted as 80% (Grimes, 2008). Information overload, where the volume of data produced has outgrown their processing and analysis capacity (Woods, Patterson, and Roth, 2002), is a growing concern for many industries, especially in the case of free-text data which traditionally relies on human oversight to extract actionable information. Henke et al. (2016) estimate that 76% of work activities require natural language understanding; therefore, developing automated methods to efficiently process natural language texts is essential.

Given the rise in the availability of digital data, it should not be surprising to discover that significant progress has been achieved within the last decade to process and analyse data of all types. Specifically, in the last 10 years, huge advances in Natural Language Processing have opened up the possibilities when dealing with natural language. Alongside the evolution of advanced machine learning methods, such as deep neural networks, these developments have caused many to say that we are currently experiencing the Fourth Industrial Revolution, an era of technological advancement which will fundamentally transform our current status quo.

These two points, the increase in digitally available construction failure reports and rapid development of methods to deal with natural language data, mean that this area is ripe for research and innovation in construction. Technological innovations are traditionally slow to filter to the construction industry. In 2016, McKinsey Global Institute analysed 22 different industries and found that construction was the second least digitally advanced in terms of digital assets, usage and labour (Gandhi, Khanna, and Ramaswamy, 2016).

Having said this, there has been increasing application of advanced analytics into the construction domain. Specific to the analysis of written documents, since beginning my research investigation in 2016, an increasing number of academic papers showcase several natural language tasks performed in the construction sector, including classification and retrieval of documents, for example Goh and Ubeynarayana (2017), Zou, Kiviniemi, and Jones (2017), Chokor et al. (2016), Marzouk and Enaba (2019), and Zhang et al. (2019).

This research examines learning from failures on construction sites in the UK. By integrating insights gained through application of modern data science and text processing methods into organisational learning processes, this research investigates reducing repetitive failures on construction sites. This research aims not just create economic advantage, by saving time and money, but is part of our moral imperative to safeguard construction employees and the public. The methodology developed also anonymises the failure data to facilitate industry-wide sharing, analysis and learning. By implementing learning processes which facilitate continuous iterative improvements, driven by failure data, it may be possible to also contribute to the prevention of catastrophic failures.

1.2 Problem Statement and Research Questions

Repetitive failures have a significant negative impact on the delivery of construction projects. The aim of this project is to investigate how to learn from these failures, especially considering novel data science methods available.

The research questions for this research are listed below. These research questions are not those from the beginning of this research journey, but rather the result of a refinement and confirmation process after critical examination of relevant literature. The initial question framing and refinement process are both captured in Section 3.2. This research focuses on RQ3 (**in bold**).

1. How does the construction industry currently learn from failure?
    - Literature search exploring learning from failure in the construction industry, comparing to norms and best practice in other industries
    - Undertake a qualitative investigation
2. What recent AI and data science methods have been used in the construction industry, and what other methods exist?
    - Literature search
3. **Which Natural Language Processing (NLP) + Machine Learning (ML) model best facilitates knowledge discovery from text-based failure data?**
    - Collect text-based failure data
    - Identify methods to convert unstructured failure data into structured forms
    - Test the accuracy of different NLP + ML models for text-based failure data
    - Identify/develop suitable knowledge discovery models
4. How is best to implement this type of learning into systematic processes for the construction industry?
    - Consolidate and discuss these findings in relation to application for the construction industry

1.3 Thesis Structure and Original Contributions

A pragmatic methodology is taken for this research. As elaborated upon in Chapter 3, by applying this pragmatic view to the research questions, the most suitable method to adopt was a multi-method research approach. This thesis is structured as follows.

A critical review of relevant literature and context is undertaken in Chapter 2, addressing RQ2 and part of RQ1. Shortcomings in the literature currently concerning RQ1 meant that further investigation was required, as outlined below. Next, the overall methodology is discussed and the research questions are revisited in Chapter 3.

The first method, a thematic analysis of learning from failure in construction, is then presented in Chapter 4. Addressing RQ1, this chapter includes specific methodological and method considerations for the qualitative investigation. Data collected via thematic analysis of semi-structured interviews is discussed in relation to literature. At the end of the chapter, the implications of this investigation upon the next method are outlined.

In particular, it is found that a significant amount of useful information on failures is captured in free-text format, currently inaccessible to many digital analytic methods. Additionally, embedding trust and explainability into AI (artificial intelligence) methods to analyse such data is vital for developing methods to learn from failures. A two-step method is undertaken to facilitate learning from failure: structuring the text data using attribute sets and knowledge discovery using these attribute sets. This tackles barriers to moving from raw data to intelligence by allowing automatic analysis of text-based failure data on construction projects.

An analogy for this process is the act of representing a landscape as a map. An aerial photo represents a huge amount of the information in the landscape, however, is too detailed and complex to easily interpret for the purpose of planning a journey. On the other hand, a line sketch containing the key features - stream/road routes and locations of buildings - is an abstract representation of the landscape, containing less of the information but is more useful to the journey-maker. In this way, refining the unstructured information contained within the text descriptions into a set of attributes reduces the complexity of the representation and allows analysis and interpretation.

Figure 1.2 presents a proposed flow for the conversion of text-based failure data into intelligence valuable for learning within industry. This illustrates the stages of data transformation and the proposed methods used (under arrows). This builds on learning from failure theory already presented to create a deliberately simple framework, which is comparable to those already implemented in other industries.

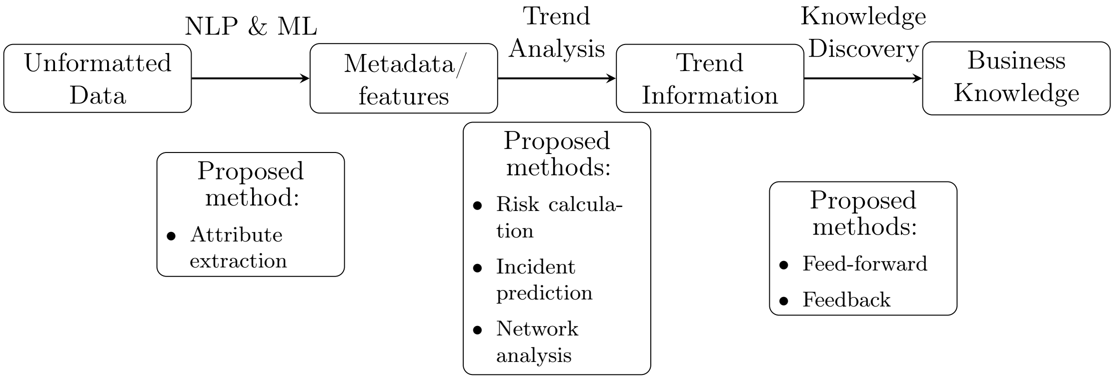


Figure 1.2: Framework

Chapter 5 encases the heart of this research. Addressing the challenge of structuring the free-text failure data (objectives 1-3 of RQ3), this chapter explores natural language processing (NLP) and machine learning (ML) methods. The assumptions for method selection and data bias are heavily reliant upon the results of the previous investigation. This is a quantitative investigation which applies Natural Language Processing (NLP) and machine learning (ML) techniques to text-based failure reports from construction projects to structure and analyse the data.

Chapter 6 outlines three possible knowledge discovery methods to apply the results of structuring this data to knowledge discovery tasks. This covers the final objective of RQ3. These methods translate the structured data into usable knowledge for inclusion in organisational learning in the construction industry.

Chapter 7 contains discussion of the results of both quantitative method steps, unstructured-to-structured data and structured data-to-knowledge as well as using qualitative information from the thematic analysis method to assess usefulness and form recommendations for application of the data analysis methods employed. This addresses RQ4.

The thesis is concluded in Chapter 8 with impact into industry and recommendations for further work. These recommendations include suggested routes for development in industry and avenues for further research.

Original Contribution

To my knowledge, this is the first multi-method investigation into the use of text-based failure data for learning in the construction industry. Recent literature, as seen later, has started to explore application of Natural Language Processing (NLP) to text-based failure data; however, they do not investigate the context and assumptions.

The first original contribution of this research is a greater depth to the understanding of the current state of learning from failure in construction.

The second contribution is development of a set of representation attributes for safety data in the UK. This used original data from a large UK construction company.

The third contribution is development of a Natural Language Processing (NLP) + Machine Learning (ML) pipeline using human-centre machine learning principles which can be trained to automatically extract attributes from text-based failure data in order to structure these data for further analysis.

The fourth contribution is a detailed cross-examination of the principles of this methodology and principles of general application of AI in construction.

Finally, this research contributes substantial recommendations for application of the findings into industry and avenues for future research.

Stylistic comments

During this research, I have written specific steps in the first-person to narrate the process. This is a stylistic choice and also reflects a philosophical decision to highlight the subjective and active role I had as the researcher in parts of this work. In particular, the discussion of methodological choice is written is the first person to highlight the importance of researcher bias as well as discussion surrounding personal reflections on the interview process detailed in Chapter 4.

1.4 Published Academic Works

The following academic papers have been produced throughout this research:

- Baker H., Hallowell M., Tixier A. J.-P. (2020) AI-based Prediction of Independent Construction Safety Outcomes from Universal Attributes, Automation in Construction, October, 103146.

- Baker H., Hallowell M., Tixier A. J.-P. (2020) Automatically Learning Construction Injury Precursors from Text, Automation in Construction, October, 103145.

- Baker H., Smith S.D., Hallowell M., Oswald D. (2020) Exploring the Association Among Dimensions of Safety Climate and Learning Organization Climate, In: Construction Research Congress, ASCE.

- Velikova M., Baker H., Smith S.D. (2018) The Meaning of Failure: Establishing a Taxonomy of Failure in the Construction Industry to Improve Organisational Learning, In: 34th Annual ARCOM Conference, 16-25.

- Baker H., Smith S.D., Velikova M., Masterton G., Hewlett B. (2018). Learning From Failure: Processes and Attitudes in the Construction Industry, In: 34th Annual ARCOM Conference (Working Paper).

- Baker H., Smith S. D., Masterton G., Hewlett B. (2018). Failures in Construction: Learning from Everyday Forensic Engineering, In: Forensic Engineering 2018: Forging Forensic Frontiers, 648-658.

Chapter 2

Literature and Industry Context

"*Mystification is simple; clarity is the hardest thing of all.* "

In "Flaubert's Parrot" by Julian Barnes

Sometimes thinking about a task or subject from a different angle gives you clarity about the crux of the matter. This is how I found writing this section. The multidisciplinary nature of my topic relied on context from many different academic disciplines, which differed in content style and philosophical approach. Many of the subjects were so entangled that I was attempting to unravel different threads which would form a picture only if interwoven. However, by stepping back to consider what my true purpose of this section was, I decided on a single word: 'clarity'. This section is not just about regurgitating past research, but also about providing my readers with clarity about the current situation so that my research purpose is perfectly clear. With this in mind, I found writing much simpler as, instead of asking 'what does the reader need to know?', I was asking 'how do I situate and make clear my purpose to the reader?' In this way, the structure for this section emerges. I lead the reader through the background knowledge and literature using a series of questions as headings, which direct the purpose of each sub-section.

2.1 Introduction

Contained in this section is a critical exploration of relevant literature and context for this research project. An essential first step in any literature review is identification of key terms and concepts. From here, exploration of these topics highlight gaps in the literature which form further research questions. This sets the logic for this section, as demonstrated in Figure 2.1, where defining key terms from the overall aim and initial research question (RQ1) leads to the identification of the research topic. It should be noted that the phrasing of these RQs match that initially used at the beginning of this research, before re-framing as seen in Section 3.2.

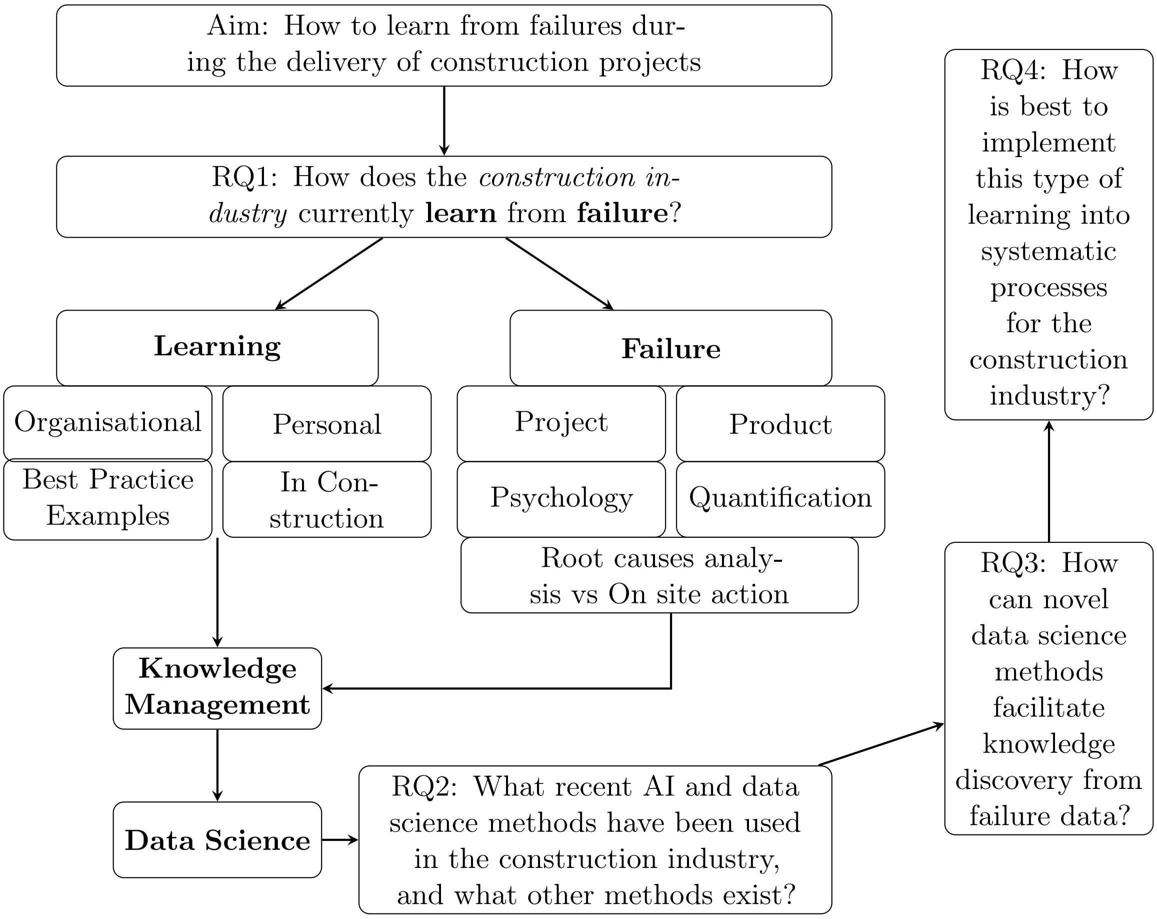

Aim: How to learn from failures during the delivery of construction projects
RQ1: How does the *construction industry* currently **learn** from **failure**?
Learning
Failure
Organisational
Personal
Best Practice Examples
In Construction
Project
Product
Psychology
Quantification
Root causes analysis vs On site action
Knowledge Management
Data Science
RQ2: What recent AI and data science methods have been used in the construction industry, and what other methods exist?
RQ3: How can novel data science methods facilitate knowledge discovery from failure data?
RQ4: How is best to implement this type of learning into systematic processes for the construction industry?

Figure 2.1: Mind Map for Concept Relation

Sections 2.2 and 2.3 target the first research question: "How does the construction industry currently learn from failure?" The first essential concept is the definition of failure in construction. Section 2.2 contextualises the notion of failure in the construction context. The next section builds on this definition to focus in on how the construction industry currently learns from failure, and compares these processes to theory and other industries, especially those known as 'learning organisations'.

Within this literature, it emerges that 'Knowledge Management' (KM) is understood to enable learning from failure in many businesses. Knowledge management and discovery describes an area of research and practice concerned with developing business acumen via robust and insightful information systems. This includes storage, access and routine analysis of data collected or collated by the organisation. This concept is explored Section 2.4.

Having identified an increasing inclusion of data science principles within knowledge management systems, Section 2.5 introduces data science and defines AI and ML. This leads to the need to also explore the context of AI and ML application in the construction industry. Research RQ2: "What recent AI and data science methods have been used in the construction industry,and what other methods exist?". Although an appreciation of these fields is given, specific consideration of ML definitions and algorithm theory is outlined later in the thesis, in Chapter 5.

Having established the 'state-of-play', this research can then move onto addressing RQ3 & RQ4 in Chapters 5 to 7.

2.2 Failure in the construction industry

At the core of this research project is a deep appreciation for the concept of failure in construction and its impact on people and processes. This subsection first examines the different ways in which ‘failure’ has been defined and then explores the challenges faced with quantifying failure, i.e. quantifying a negative/something which hasn’t happened. Another important aspect of failure is the psychological reactions and attitudes these provoke. The literature surrounding this is contextualised for the construction industry. The final section is a philosophical discussion exploring root cause vs immediate cause.

2.2.1 What is the definition of failure?

Failure, as well as being a provocative subject, is not easily defined. Even the clearest dictionary entry on ‘failure’, given by the Cambridge dictionary, splits its definition of the noun into three distinct uses:

- The fact of someone or something not succeeding, for example, “The meeting was a failure” (= the meeting was unsuccessful in its aims).
- The fact of not doing something that you must do or are expected to do, for example, “There are serious penalties for failure to comply with the regulations” (= there are serious penalties for not complying with the regulations.)
- The fact of something not working, or stopping working as well as it should, for example, “The number of business failures rose steeply last year.” (= the number of businesses which ceased to work rose steeply last year.)

These different, and yet conflated, definitions of failure make discussion about failure difficult. Even the individual definitions bring with them a host of problems. Consider the first definition: “the fact of someone or something not succeeding”. In order for this definition to be useful, the factors which determine ‘succeeding’ must be **known**. In the second, which states failure is “the fact of not doing something that you must do or are expected to do”, it must be **known** what you are expected to do so that it is **known** when deviation occurs. Equally, in the final definition, it must be **known** what something looks like when it is working to be able to usefully define failure as “something not working”. A key point in defining failure appears to be in **knowing** what stipulates the opposite. It is useful, therefore, to consider what is meant by ‘successful’, ‘expected’ or ‘working’ for construction projects.

This sub-section considers each of these three concepts in the context of construction projects. The literature found on ‘failure’ in construction is then presented. As will be seen, this is significantly smaller than that exploring success.

What does *success* mean in construction?

Project success can be split into two key concepts: project management success and product success (Baccarini, 1999). This distinction can also be applied to construction projects.

Project management success

A large body of research concerning failure in construction revolves around project failure. Coming from the project management discipline, this academic sphere is vast and has been the topic of entire careers’ worth of work.

Starting from a simplistic ‘textbook’ view, project success is often understood as achieving a specified quality of product in a defined time frame and cost. This is often discussed in terms of the ‘Iron Triangle’, also referred to as the ‘Triple Constraint’, a trade-off between cost, quality and time, as seen in Figure 2.2. This simple illustration facilitates easy conversation about the trade-off between these success criteria. Despite being a popular model since its inception in the 1970s, there has been wide criticism of its applicability.

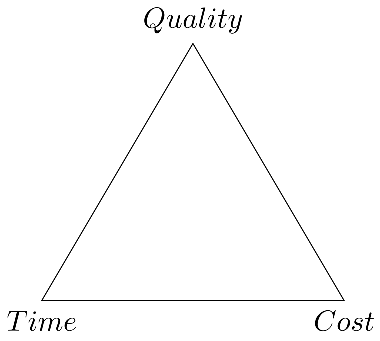


Figure 2.2: Iron Triangle

In essence, much of the criticism of the Iron Triangle model revolves around the manner in which it engages management professionals. Hoorn and Whitty (2015) frame this issue as a philosophical disconnect between academic concepts of project management and the lived experience. They argue that the positivist-heavy Cartesian philosophy dominant in project management conflicts with the lived experience of projects. A key aspect of Cartesian or dualism philosophy is that “human beings are separate from discrete objects in the universe” while Hoorn and Whitty (2015) propose that a more suitable view of project management is a network where human beings, objects and phenomena interact. This aligns with their proposed philosophical lens using the ontological perspective presented in Heidegger’s Being and Time. Van Der Hoorn and Whitty (2015) present the Iron Triangle as a management tool which “veils the complexity of the project environment and managing the significant number of interrelated factors” by presenting the project management experience as a linear trade-off between time, cost and quality. To this end, critics of the Iron Triangle worry that naive application of this tool could lead to a blinkered vision of project management and overemphasise management of these three criteria, to the detriment of other interrelating factors.

It is therefore beneficial to separate ‘success criteria’ and ‘success factors’. Success criteria are outcomes which count towards the success of the project. Meanwhile, success factors facilitate the achievement of these outcomes. Rather than viewing the Iron Triangle as a project management tool, it could be considered an illustration of success criteria, which then engages the professional to consider the factors which manage each criterion. However, this still emphasises the importance of cost, time and quality as success criteria. So first it should be considered whether ‘time, cost and quality’ are the best, or only, success criteria to consider before exploring success factors later in the section.

A recent review by Pollack, Helm, and Adler (2018) catalogues the ways in which the Iron Triangle has evolved over time. They document that a large number of authors turned the debate to the definition of the criteria on the vertices, in particular redefinition of the ‘quality’ vertex. Chan, Scott, and Lam (2002) identifies ‘quality’ as a subjective measure, while time and cost are objective. This sparks debate on its suitability. ‘Scope’, ‘Performance’ and ‘Requirements’ are all suggested by multiple authors as more suitable alternatives to ‘Quality’, however, the conclusion of Pollack, Helm, and Adler (2018) is that the relevance of these criteria definitions lie within the context. It appears that the re-definition of these vertices revolves more around assessment of the success criteria rather than radical change in the criteria itself.

This conclusion, like that of Hoorn and Whitty (2015), indicates that engagement with the Iron Triangle model and critical thought are required rather than blind application of a set norm. For example, in the construction industry, 'quality' is almost synonymous with 'compliance to engineering standards and contractual specifications'. This closed definition, bordering on an objective definition, makes 'quality' a suitable proposed third vertex for construction.

Other authors have made up for the shortfalls of the Iron Triangle model by adding other criterion, either in the form of additional vertices or inclusion of the Iron Triangle in a larger framework. Both Atkinson (1999) and Shenhar et al. (2001) propose that the Iron Triangle as one of four criteria which should be considered in defining project success. While they both suggest that the second and third criteria should be the benefits to the organisation and to the stakeholders, Atkinson (1999) puts forward 'Information Systems' as a fourth dimension while Shenhar et al. (2001) market this dimension as 'Preparing for the Future'. However, both of these dimensions deal with improving the organisation's processes and ensuring lessons learnt during the course of the project are retained to benefit future projects.

An important property of these additional dimensions, proposed by Atkinson (1999) and Shenhar et al. (2001), is the time periods of the project being assessed. While the 'Iron Triangle' still forms the core criteria for the delivery phase, the other criteria deal with medium to long-term success of the project. Chan, Scott, and Lam (2002) also group their success criteria into three phases: pre-, post- and during construction. When identifying success criteria for construction projects (or any project), different stakeholders have different timescales which they prioritise for assessment. For example, a contractor may base their definition on the construction phase itself, with criteria like their own profit margins, safety of their employees and achieving contractual milestones. On the other hand, construction clients may be more concerned with total cost, handover dates and compliance. Meanwhile, for the eventual asset owners, they may not be interested in actions during the construction phase but the legacy it leaves: Is the build to a high quality? Will maintenance costs be low? Does it achieve its purpose? The majority of these criteria could be discussed under the heading of time, cost or quality. Again, supporting the notion that the 'Iron Triangle' criteria are valuable to spark critical discussion.

Arguably, for the construction industry, a significant addition for the delivery phase to the three general criteria is the inclusion of safety. Chan, Scott, and Lam (2002) found that 'Health and Safety' was the fifth most prominent success criterion in literature between 2000-2010 after Iron Triangle criteria and stakeholder satisfaction. They classify Health and Safety as an objective measure. This was suitable, at the time, with the prominence of lagging statistics for safety, such as Accident Frequency Rates (AFRs). However, since then, assessment of subjective H&S factors has developed, such as safety culture assessments. This is discussed further in the quantifying success subsection (2.2.2).

As mentioned, 'stakeholder satisfaction' is often quoted as an important success criterion (Chan, Scott, and Lam, 2002). This criterion is postulated as a subjective measure of opinions throughout the projects, with stakeholders including clients, construction staff and ultimate users (i.e. the public for infrastructure projects). This criterion can be applied to the project management success or the product success.

Despite the criticisms held against it, the Iron Triangle is still a widely used metric in practice for the construction phase. Pollack, Helm, and Adler (2018) reported that the majority of project managers made use of this model. The sustained prevalence of this model indicates the importance of these three criteria - money, time, quality - to the success of the project. This research, therefore, accepts the Iron Triangle as a valuable thought tool for exploring success criteria in the construction industry. Safety and stakeholder satisfaction are also relevant success criteria for the construction industry.

Product success

The product of a construction project should last far beyond the time frame of the project delivery. Aspects of product success have been captured in previously mentioned success criteria within the post-delivery phase, such as the stakeholder benefits covered in Atkinson (1999).

Three main modes of construction product success are discussed. These are structural, functional and stakeholder satisfaction.

In construction products, structural integrity is a key factor in product success. The Code of Hammurabi (Approx. 1754BC) is a historically significant text from the reign of the Babylonian King Hammurabi (1792-1750 BC) consisting of 282 laws which are the "the most complete legal compendium of Antiquity, dating back to earlier than the Biblical laws" (Claire, 2009). It documents the first known legal repercussions of structural failure for construction workers. Those laws pertaining to the built environment read:

> 229. If a builder has built a house for a man and has not made his work strong enough and the house he has made has collapsed and caused the death of the owner of the house, that builder shall be killed.
>
> 230. If it has caused the death of a son of the owner of the house, they shall kill that builder's son.
>
> 231. If it cause the death of a slave of the owner of the house, he shall give a slave for the slave of the owner of the house.
>
> 232. If he has destroyed possessions, he shall make recompense whatever he destroyed. Moreover, since the house he had built collapsed because he had not made it strong enough, he shall rebuild the house which collapsed from his own resources.
>
> 233. If a builder has made a house for a man and has not made it strong enough and a wall has toppled, that builder shall strengthen that wall from his own resources.
>
> Note: This translation is from Richardson (2004). Other translations exist and differ slightly in terminology, however, the core semantics remain consistent.

While modern law has, in general, moved on from this 'eye-for-an-eye' approach to legal punishment (also known as 'laws of retaliation' or 'lex talionis'), structural integrity remains a key aspect in the success of a construction project with legal repercussions for failure of this type. While on some level this may be trivial, i.e. 'if it hasn't fallen down, it has succeeded', there is a whole host of nuance in assessment of a whether a structure has failed. The American Society of Civil Engineers (ASCE) defines forensic engineering as "the application of engineering principles to the investigation of failures or other performance problems". This covers investigation of structural failure and collapse, as well as other technical failures. It is perhaps better to classify this success criterion as structural integrity and technical resilience.

The product must also be functional. Chan, Scott, and Lam (2002) define this as "the degree of conformance to all technical performance specifications", however, some construction products fail to be 'fit-for-purpose' due to incorrect solution identification. This means that the project may be successful in meeting the defined scope or technical specification but this specification fails to meet the desired function. Additionally, for construction projects, the achievement of this criteria could change over time. What is considered a successful project at one point may no longer be so in several years. The infrastructure needs of today are different than those in 5, 10 or 50 years time. A product which meets those needs today may not be successful at meeting the future needs in several years.

As previously mentioned, stakeholder satisfaction is widely acknowledged as an important success criteria, both in assessing success of the project management and the product.

Overall success

Here, proposed project management success criteria include: money, time, quality and safety. Meanwhile, product success revolves around functionality and structural stability. More subjective criteria for success revolve around stakeholder satisfaction. However, it is clear that there is no consensus on a finite set of criteria for success and any framework should be used to structure critical thought rather than as an objective premise.

There is also a significant gap in literature concerning what construction professionals consider failure criteria, as opposed to success criteria. While discussion so far defines failure as the opposite of success, it does not follow that the most important success criteria are also the most important criteria to avoid failing in. To address this gap, a qualitative investigation is presented in Chapter 4.

A story of success

The A14 improvement project to improve the road between Cambridge and Huntington is perhaps the most recent example of a UK infrastructure success story. The £1.5bn project aimed to improve this congested route by widening and improving junctions and road sections as well as building a large new 12 mile by-pass. The scheme opened to public traffic in May 2020, 8 months ahead of schedule, with one key section having been opened since December 2019. With only auxiliary works still to complete, this early completion has been heralded as a huge success.

Moreover, the project is on budget and appears to have achieved high satisfaction from the client, Highways England, who has expressed pleasure with both the fast-tracked delivery and awards the project has been involved with. The project was the first to be awarded 'Ultra Site' status, the highest award from the Considerate Constructors Scheme (a not-for-profit, independent organisation founded in 1997 to raise standards in the construction industry) for their outstanding standards. It was also one of five nominees for BBC Countryfile Magazine's 'Conservation Success of the Year' category in 2018. (*A14 Integrated Delivery Team – A14 Cambridge to Huntingdon Improvement Scheme _ ccscheme* 2019)

Much of this success has been attributed to their 'integrated' team delivery. The joint venture between Costain, Skanska and Balfour Beatty used an innovative approach to project management to encourage collaborative working at all levels of the supply chain. (*A14 Improvement Scheme Progress* 2020)

Of the key success criteria identified for construction projects, the only one conspicuously absent from the news articles is product quality. As mentioned, this is often defined as 'fit-for-purpose' which not only relies on correct solution identification but also changes depending on time period and stakeholder perspective. Concerning this project, solution selection was conceived over a five-year period from 2011-2016. While the selected solution has been delivered, only time will tell whether this criterion is fulfilled.

However, as proof that success is assessed differently for different stakeholders, a number of the public are still unsatisfied. One member of the public reported that the adjustments made to a junction on his daily commute produced a bottleneck where none was there before. Meanwhile, during the construction, a number of complaints were made about the noise which prompted the construction team to adjust their protective equipment.

Ultimately, this example demonstrates the use of success criteria in popular media to assess the success of an infrastructure project. The criteria identified in the news and public domain reflect those identified in the project success literature.

What does *expected* mean in construction?

Another dictionary definition of failure in the Cambridge dictionary is the 'fact of not doing something you ... are expected to do'. However, expectation is subjective, intangible and culture dependant. Expectation in the workplace stems from shared values and priorities. To appreciate the different expectations which may be present around the globe, it is interesting to observe the different priorities construction professionals place upon success factors or criteria. As mentioned earlier, success factors and success criteria are different concepts. While 'criteria' indicate the outcomes, 'factors' are those steps which facilitate the achievement of these outcomes.

There has been a cluster of research exploring success factors in different cultural contexts. The methodologies for these studies have been generally consistent, with identification of proposed success factors in literature then undertaking a quantitative questionnaire collecting opinion data from construction professionals to identify which factors they believe are critical in achieving success. In surveying professionals in this way, the authors have not separated out different success criteria and so have aggregated all types of project and product success under one umbrella. In other words, by asking 'what is important for success?' rather than 'what is important for financial success or scheduling success etc?', these surveys also require the participants to define success themselves. As such, these surveys also hold bias as to the success priorities of the professionals themselves.

Osei-Kyei and Chan (2017) is the only paper found directly comparing the priorities of two countries using the exact same questionnaire. They only compare 15 factors while some others compare up to 39, e.g. (Toor and Ogunlana, 2008). Additionally, their concept of 'factors' does not line up with the definition held by other papers. They also include 'success criteria' such as 'profitability'; however, this study can be used to illustrate the significant differences in the priorities of the construction professionals in different countries. The most significant rank difference is the 'profitability' of the project, ranked 1st for Ghanaian construction but 11th for Hong Kong. The authors attribute this difference to the differences in risk profile and financial pressures experienced in each country. It is clear that the different cultural undertones and pressures affect the priority given to different success criteria.

This study highlights a significant limitation to this cluster of research. In interpreting opinion based questionnaires, there is a lack of literature separating the prioritisation of success criteria from success outcomes. In other words, it is unknown whether a factor is valued because the construction professionals prioritise the success criteria it is most associated with, or because they value that factor in general. As an example, if a construction professional ranks 'strong projects finances' highly as a success factor, it is unknown whether this is because she also ranks 'money' highly as a success criterion, or that she appreciates that this factor facilities many types of success. In further discussion of these studies, it is necessary to appreciate these biases which the questionnaires hold.

Table 2.1 shows a comparison of the top success factors for a selection of surveys, as well as including the top success criteria from the previous discussed study by Osei-Kyei and Chan (2017). These quantitative surveys aimed to assess the priorities of construction professionals in regards to success factors; factors about the project which they believed helped achieve project success. In comparing the factors identified around the globe, it is clear that expectations change from country to country. However, there are also quite a few similarities. Most of the surveys found that construction personnel value clear and appropriate contractual agreements, which define scope and risk allocation. Additionally, the competence of different personnel ranked highly for several countries.

Project financing and resource was also a prominent theme within the success factors. This could indicate a global prioritisation of money as a success criteria. However, it could also

Table 2.1: Comparison of Critical Success Factors in different countries

Context	Success Factors	Reference
UK (PPP/PFI)	1. Strong private consortium 2. Appropriate risk allocation 3. Available financial market	Li et al. (2005)
Thailand	1. Effective project planning and control 2. Sufficient resources 3. Clear and detailed written contract 4. Clearly defined goals and priorities of all stakeholders 5. Competent project manager	Toor and Ogunlana (2008)
India	1. Awareness and compliance with rules and regulations 2. Effective partnering among project participants 3. Pre-project planning and clarity in scope	Tabish and Jha (2011)
Pakistan	1. Decision making effectiveness 2. Project managers experience 3. Contractor's cash flow	Saqib, Farooqui, and Lodi (2010)
Malaysia	1. Contractor's competence and experience 2. Project financing 3. Team leader's competence	Yong and Mustaffa (2013)
Context	Success Criteria	Reference
Ghana (PPP)	1. Profitability 2. Meeting output specifications 3. Adherence to budget 4. Adherence to time 5. Reliable and quality service operations	Osei-Kyei and Chan (2017)
Hong Kong (PPP)	1. Adherence to budget 2. Adherence to time 3. Effective risk management 4. Meeting output specifications 5. Reliable and quality service operations	Osei-Kyei and Chan (2017)

reflect on the competitive nature of tendering processes and the effects that inadequate resource availability can have on the morale and ability of professionals produce the project. As previously mentioned, 'overspend' on projects can be due to an inaccurate initial budget or can be due to excessively low bidding. Low bidding can result from overly competitive tendering, which is often exacerbated when the client organisation is under financial pressures.

Pertinent to this research topic, Toor and Ogunlana (2008) found that of the 39 factors participants were asked about, the one with the lowest score was 'using up-to-date technology and automation for construction work'. It could be postulated that this is due to the emerging nature of Thailand's economy, however, a study by McKinsey Institute in the USA also found that the construction industry is the second lowest for digitisation of the 22 industries they assessed (Gandhi, Khanna, and Ramaswamy, 2016). While limited in scope, these studies spark inquiries into whether uptake of modern automation and digital technology in construction is hindered by human factors - such as lack of priority - rather than lack of applicability. This could be (and is - see work by fellow Costain PhD Carolina Toczycka) an entire research project in its own right. However, the salient point for this research is that any new digital solutions for the construction industry will have to overcome significant barriers and should, therefore, be developed with full cognizance of the human factors specific to the problem statement.

These studies show that, while there are differences in expectations around the globe, some re-occurring themes hold strong. These include personnel competence, strong project governance and project finance. It is apparent from this literature that it is necessary to consider success (and failure) within the cultural context and allow this context to guide research decisions.

What does *working* mean in construction?

This research considers that 'working' in construction refers to physical objects such as equipment, in order to avoid conflating this type of failure with 'the opposite of success'. This helps to facilitate clear discussion and was

The dictionary definition is clear that this type of failure is where a system or object which was performing a function then stops working. This definition therefore depends on the identifying when something is working, and when this ceases. This is then only as complex as the system or object which is being considered. For simple and well-defined objects, this is perhaps the easiest definition. For physical objects, 'working' in construction could be defined in terms of the equipment being functional.

However, for more abstract or complex systems, such as management systems, the issue lies in identifying the whole system, when it is working and when it has ceased. For example, when assessing if a safe system of work is 'working' in construction, the project engineer may say that it has ceased to work when an incident occurred, however, how is she to know that it was working beforehand and that workers were not simply lucky? In this case, it is simpler to sort this type of system into the first definition of failure, as a lack of success, and say that the system is a success when nobody is injured. But this encourages, once again, the Cartesian philosophy that human beings are separate from the objects and processes they interact with. Reflecting on the arguments set out by Van Der Hoorn and Whitty (2015) about the philosophy of project management, the authors of that article would most probably encourage the complex systems view encouraged by this definition, rather than the success criteria view set out previously in defining 'success'.

What do these definitions mean about failure in construction?

To this point, discussion has been focused on success and defining failure as the opposite. It is implied that by achieving success, it is possible to avoid failure. Literature on success splits into defining success criteria - desired outcomes - and success factors - steps or traits which help achieve these desired outcomes. As will be seen in the next sub-section, both of these aspects are assessed throughout the project life-cycle using performance assessments, where managers and key stakeholders have identified what they believe to be important to the success of the project and decided a method to measure those aspects. However, does it actually follow that the same criteria and factors which are important for success are also important to avoid failure?

While literature defining failure in construction is much sparser than project success literature, there is still a collection of studies investigating failure in construction projects. These studies can, in general, be grouped into two classes: (1) case studies of construction project failures and (2) questionnaire or group-sourced data concerning failure causes.

The first type of failure literature follow root cause analysis, as introduced later in Section 2.2.4. These papers or reports document one or several construction or infrastructure failure(s) and analyse their causes - which can be referred to as ‘failure factors’. These case studies tend to be infamous in their own right, having gained notoriety in the news for their failure. While important learning can be gained from these studies, their generalisability is extremely limited as they are based entirely on a single instance which, by its very nature as an extreme event, is an outlier. This is discussed further in Section 2.2.4.

The second type of failure literature are those using questionnaire or other group-sources of data, such as interviews or focus groups, to analyse the suspected traits or causes which lead to failure. In similar ilk to those investigating success (see previously in ‘What does *expected* mean in construction?’), these are limited by a lack of explicit definition of what failure is, before launching into the factors which cause it. Pinto and Mantel (1990) identified 30 years ago that “the critical factors associated with failure depended on the way in which failure is defined, [this] suggests that it is necessary to know considerably more about how project managers define failure (and success) and, indeed how the parent organization makes judgments on the matter.” Despite this, authors still often fail to adequately define failure in their work, or make clear that they have adequately communicated to their participants which definition they have adopted. In those which do define the failure criteria, such as Al-Zwainy, Mohammed, and Varouqa (2018), the failure criteria are limited to time and cost overrun. These studies are also limited in their response rates, with limited participants to each study.

In conclusion, there is a lack of agreement and foundation to the definition of failure in construction, especially when limiting the literature to a UK context. This is essential for this research and therefore must be investigated further before research into suitable ‘learning from failure’ systems can begin. This literature gap led to the qualitative investigation presenting in Chapter 4.

2.2.2 How is failure (or success) quantified?

Any discussion on different types of success and failure in construction could be considered limited, especially to those in management, without methods of measuring them. Quantifying success, also known as performance measurement, is a core task for any project manager to demonstrate the return on investment to various stakeholders. Investment can be any type of resource. This could simply be financial investment, however, stakeholders also invest time and effort into the project - team members want to see its success too!

The construction industry has historically tended to emphasise money, time and quality performance measures (Atkinson, 1999). In the last 20 years, these assessments have expanded to include measures of other success criteria, such as stakeholder satisfaction. Tripathi and Jha (2018) recently conducted a systematic review which identified 20 different measures in construction performance literature. They then used questionnaires to ask construction professionals in India to rank these measures in terms of importance. Table 2.2 lists these 20 measures in order with suggestions on their calculation method. While columns 1-3 are directly from the paper (Tripathi and Jha, 2018), a fourth column has been added which considers the type of success criteria these measures and calculation methods aim to assess. As seen, the majority (75%) are related to time, money or stakeholder satisfaction criteria. Even the performance measure for rework, which could be considered by engineering professionals as a measure of quality, has a calculation method which deals directly with incurred cost - not quality of build. The exceptions to this rule are Health & Safety and Environment success criteria. There are also some success factors listed - measures of a variable which affects success but is not an outcome in its own right - for example, 'wages'.

Lagging vs Leading

The first thing to note about the suggested calculation methods is that the majority of them are lagging indicators. This means they require data after the completion of the project or time period being assessed to calculate the performance. Lagging indicators have been historically dominant in management performance assessment as they are often considered direct measures of the success criteria - for example, a suggested measure of achieving the 'on time' criteria is to calculate the percentage of projects (or tasks) which occurred on time. This is assessed after the event and cannot affect the outcome.

Recently, leading measures have grown in popularity. These measures can be said to focus on assessing success factors and attempt to keep the 'finger on the pulse' of the project, allowing for adjustment throughout the project life cycle. An example from Table 2.2 is the 'wages' performance measure, which could be postulated to motivate employees and therefore be a success factor to achieve several different success criteria. This was ranked last by the professionals asked, however, it is unknown whether this is because the Indian professionals prioritised performance measures which dealt directly with assessment of success criteria, or because they deem wages to be irrelevant to the performance of a project.

An area of construction which has seen attention in leading indicator research is that of Health & Safety. As with most success criteria, lagging measures (mainly incident numbers) remain dominant and are considered essential to reporting and documenting past performance. However, leading measures which allow the project to adjust and adapt to keep people safe are extremely appealing. So far, these have focused on quantitative measures such as assessing the occurrence of proactive safety tasks. For example, Hallowell et al. (2013) summarise measures aggregated from literature such as the number of safety observations per 10000 man hours and the frequency of decision-makers safety walkthroughs on site. However, Oswald (2020) identifies the lack of assessment of the quality of these tasks - a focus on quantity over quality - as primary limitation to the current implementation of these measures. This can lead to inspiring the incorrect behaviours and additional 'tick-box' tasks for the sake of increasing performance metrics, rather than performance.

Objective vs Subjective

Many of the assessment methods are known as 'objective' measures defined as measures calculated using numerical project data and mathematical formulae. These are extremely prominent; only 4/20 examples in Table 2.2 suggest a subjective measure. Referring to these as 'objective' measures is actually misleading as their selection, as explained next, is a subjective task and the selection (or omission) of different measures can affect how well the project appears to be succeeding (or not as the case may be). The way in which the numerical data is collected can also introduce biases into this assessment. Additionally, these numerical measures yield limited insight into *why* a value is obtained.

So called 'subjective' measures are those calculated using opinion or qualitative data. These

Table 2.2: Performance Measures and Methods of Measurement Tripathi and Jha (2018)

Number	Performance Measure	Calculation Methods	Category
1	Good track record of timely completion of the projects	$\frac{\text{Number of projects delivered on or before schedule}}{\text{Total number of projects}}$	Time
2	Good relationship with client	% of repeat clients $=\frac{\text{Number of repeated clients}}{\text{Total number of clients}}$ Low dispute and litigation Timely payment from clients	SS*
3	Customer satisfaction in terms of product and services	Customer satisfaction survey	SS*
4	Client satisfaction in terms of product and services	Client satisfaction survey	SS*
5	Predictability of time in design and construction	$\frac{\text{Actual time - Anticipated time}}{\text{Anticipated time}}$	Time
6	Productivity of employees	Productivity $=\frac{\text{Works units completed during a given time period}}{\text{Associated cost in terms of man-hours or dollars}}$	Time & Money
7	Predictability of cost in design and construction	$\frac{\text{Actual cost - Anticipated cost}}{\text{Anticipated cost}}$	Money
8	Higher annual growth rate of the organisation	Return on assets (ROA) $=\frac{\text{Company's annual earnings}}{\text{Total assets}}$ Return on equity (ROE) $=\frac{\text{Net income after tax}}{\text{Share holder's equity}}$ Return on capital (ROC) $=\frac{\text{Net income -dividends}}{\text{Total capital}}$	Money
9	Cost performance of projects	$\frac{\text{Number of projects completed within tender cost}}{\text{Total number of projects}}$	Money
10	Annual construction demand/market share	$\frac{\text{Company's volume of work in the market}}{\text{Total volume of work in the market}}$	Money
11	Health and safety consciousness	Safety performance $=\frac{\text{Number of reported accidents}}{\text{Average number of employees}}$	Health & Safety
12	Optimum liquidity ratio	Current ratio $=\frac{\text{Current assets}}{\text{Current liabilities}}$	Money
13	Low staff turnover	$\frac{\text{Number of employees leaving the organisation in a year}}{\text{Average number of employees in that year}}$	SS*
14	Rework/defect rectification	Rework factor $=\frac{\text{Total cost of rework}}{\text{Total construction cost}}$	Money
15	Higher profitability ratio	Gross profit margin $=\frac{\text{Profit before tax and interest}}{\text{Total revenues}}$	Money
16	Impact on environment	Use of low natural resources Low production of waste Preservation of plants and trees etc.	Environment
17	Adopting learning and growth culture	$\frac{\text{Amount spent for learning and growth}}{\text{Turnover of the organisation}}$	Factor
18	Size of the organisation	Turnover of organisation Market share Number of employees	Factor
19	Impact on society	Low noise pollution Less disturbance to the occupants due to vehicle movement etc.	SS*
20	Higher wages	Wages of the employee with respect to the average wages in the industry	Factor

* SS = Stakeholder Satisfaction

often attempt to incorporate more intangible success criteria, like stakeholder satisfaction. Oswald (2020) argues that for every quantitative measure, there should be a complementary qualitative one which can explain why the numerical value is what is it. However, currently, these are considered too difficult to implement. The analysis required, as with lots of qualitative analysis, would be manual and time-consuming.

Choosing assessment methods

The selection of performance assessment can be considered a subjective task. It is important to choose these performance assessments carefully as using these measures incentivises to certain behaviours, both intended and unintended. For example, the necessary H&S statistics about number of reported accidents should incentivise safe behaviour to minimise the number of incidents; however, it could exacerbate natural psychological responses to failure such as avoidance and lead to lack of reporting. This is discussed further in the next section. Some measurements, such as Health and Safety and environmental impact, such as air pollutants, have to be reported by law in the UK, but how do decision-makers choose which other performance measures to use on their project?

Yang et al. (2010) performed a critical review of the performance measures adopted by the construction industry and found that the most frequently adopted frameworks for their selection are: the European Foundation for Quality Management excellence model, balanced scorecard model, and key performance indicators model. These frameworks structure the performance assessment selection to attempt to ensure a holistic assessment of a project. While many different frameworks for choosing critical performance indicators exist, Lin and Shen (2007) identified four common factors:

1. Multiperspective indicators are needed to measure performance;

2. Indicators based on characteristics of organizations or projects in different industries need to be developed;

3. Continuous measurement of performance is encouraged to achieve the best practice; and

4. Real-time feedback is necessary to make on course corrections.

Ironically, these factors conflict with the research which has already been presented about industry practice. In practice, it appears that indicators are still dominated with quantitative measures of time and money, while characteristic indicators are slow to develop. Meanwhile lagging indicators, rather than continuous leading indicators, still dominate and feedback loops are weak.

What about failure?

Again, this discussion has revolved around quantification of success, not failure. This is because, as in defining failure, the quantification of it is generally held as the opposite of success - a 'bad' performance assessment equates to a failure. Once again, an investigation is required to confirm whether this holds true or not. This is included in the preliminary investigation, presented in Chapter 4.

Interestingly, rework, safety and environment performance measures appear to be the only measures which actual 'measure' a failure. This suggests that there is something about these where we are more interested in the failure event than the success scenario.

2.2.3 How does the psychology of failure affect failure processes in construction?

It would be remiss not to mention the psychology of failure. This is essential to establishing context as it affects processes, behaviours and identification of failure.

Failure is not a pleasant experience. Despite business moguls and self-help guides stressing the importance of reflecting on our failures and embracing the learning they hold, it is hard and uncomfortable to do so even in the confines of one's own mind, let alone in a public or semi-public forum. This leads to a paradox where the desire to exploit the learning from failures conflicts with the desire to keep them hidden.

A key factor in this conflict is the degree of accountability expected of the individual. Dekker (2009) notes every failure leads to questions centering on 'whose fault?' Since failures such as Three Mile island and twin Boeing 747 disasters in the 1970s, catastrophic failure events are no longer considered 'force majeure' but organisational failures born of people and their decisions. This emphasis on accountability has consequences. Dekker (2009) nicely summarises the issue:

> *Criminalization of any act is not just about retribution and explanation of misfortune, but also about putative deterrence, and so it is with the criminalization of human error. [...] The deterrence argument is problematic, however, as threats of prosecution do not deter people from making errors, but rather from reporting them. [...] The anxiety and stress generated by such accountability adds attentional burdens and distracts from conscientious discharge of the main safety–critical task.*

In response, many safety-critical industries have created (or attempted to create) a psychological safe environment for reporting and learning from incidents. This is part of the goals of a 'just culture'. Reason (1998) explains that, at its core, a just culture requires an organisation to differentiate between acceptable and unacceptable behaviours and adjust its punitive system to account for this, not to simply assign punishment based on outcomes. For example, it should matter if an act was deliberately malicious or a genuine mistake.

One way industries implemented this is by creating confidential reporting systems, where the people submitting the report are anonymous, like those developed at NASA and by British Airways. In the construction industry in the UK, CROSS (Confidential Reporting on Structural Safety) has been run jointly by the ICE and IStructE since 2005 to capture issues of structural safety which may have otherwise been ignored. However, these systems are voluntary.

Current dialogue surrounding 'just culture' follows three strands, originating from critiques of the concept. Heraghty, Rae, and Dekker (2020) identify these as:

1. A philosophical debate surrounding the applicability of reductionist models to human error - Similar to the debate presented by Hoorn and Whitty (2015) for management models like the 'Iron Triangle', there is a body of research which recommends caution when attempting to present complex workplace dynamics as simplistic models. In particular, they advise caution against using the models outside their intended purpose, for example, Heraghty, Rae, and Dekker (2020) states that tools intended to facilitate a just culture could be used to "justify the use of retributive justice rather than to determine if it is needed at all."

2. The role of blame - There has also been a rise in associated discussion about 'no blame culture'. Reason (1998) stated that a 'no blame' culture is neither feasible or desirable. However, he was defining 'no blame' as "a blanket amnesty on all types of unsafe behaviour". However in more recent debate authors, such as Dekker (2009), argue that 'no blame' is not the same as 'no accountability'. They advocate 'blame free' systems as a way of increasing accountability as it increases reporting and openness. Other authors have turned their attention to the psychological harm blame can cause, citing this as an independent reason to minimise blame in failure investigation (Heraghty, Rae, and Dekker, 2020).

3. Practical application of 'just culture' - There is also a body of research exploring the use of the 'just culture' theory in practical management and legal proceedings. Recently, Heraghty, Rae, and Dekker (2020) reported three key findings in real application of 'just culture' to incident cases: (1) "Clear rules in the policy conceal fuzzy and subjective decision-making in reality"; (2) "Language shapes and influences every part of the process and out-comes"; and

(3) “Accident analyses cannot be treated as stand-alone “fair” processes, separate from the relationships before and after”.

Currently in construction, there appears to be a lack of adoption of ‘just culture’ methodology. Oswald et al. (2018) describes the current situation for Health and Safety as a ‘compensation culture’. They note that a compensation culture creates “a belief that when accidents do occur someone must be at fault”. This is at odds with the theory of a ‘just culture’, and the authors found that this caused fear and lack of ownership of failures on site, resulting in “blame and a reduction in safety learning opportunities; excessive paperwork in a safety management system; and a lack of worker engagement”. This was compounded when the possibility of fraudulent claims were also considered, creating an atmosphere of distrust. While this ethnographic study is limited to a single construction project, the implications are not. In designing any system which interacts with failure on site, it is essential to be cognizant of this existing culture surrounding failure and anticipate systematic barriers it causes.

It is clear that any organisational system dealing with failure must acknowledge and anticipate the cultural impacts it will have, as well as anticipate the barriers human factors will present. However, there is limited literature examining failure in construction with this light. This is addressed, along with the previously identified gaps regarding failure, in the qualitative investigation in Chapter 4.

Comparing two industries: Aviation and Healthcare

"We don't do investigations." This was the response in 2005 when Martin Bromiley approach the head of the Intensive Care Unit requesting an investigation into his wife's death following a routine operation. Syed (2015) documents the tragic story in his book, 'Black Box Thinking', which illustrates deeply ingrained attitudes held in the medical field towards failure. These attitudes are grounded in evasion and self-justification. Medical mistakes are explained away as 'one of those things'. The medical professionals are not deliberately avoiding blame or learning; they are simply products of the culture in which they operate.

This attitude is often compared to that of aviation. Often heralded as a pioneer for learning and just culture, aviation has a long history of adopting systems and post-accident investigations which facilitate learning. They achieve this through data transparency and by openly communicating these lessons. When professionals can openly appreciate the value in error reporting and are not penalised for doing so, a just culture is born. It should be noted that there have been notable exceptions to this cultural norm in recent years, such as the problems surrounding Rolls-Royce's Trent 1000 engines used to power Boeing's 787. Aviation should be vigilant that the 'compensation culture' prevalent elsewhere does not degrade their historically strong systems.

So why do the cultures appear so different in these two industries? Perhaps the first item to address is the nature and timings of the 'trigger events' which motivated each of these industries to adopt these practices in the first place. In aviation, the widespread adoption of learning from failure and just culture is often attributed to large aviation disasters in the 70s. United Airlines flight 173 in 1978 is often acknowledged as a watershed moment in aviation safety (Syed, 2015). These news-worthy events meant that something 'must' be done and done publicly across the industry. In comparison, Boysen (2013) attribute the rise in interest in the medical profession to the publication of the Institute of Medicine (IOM) report, To Err is Human, in 2000. This call to arms "sets forth a national agenda [...] for reducing medical errors and improving patient safety through the design of a safer health system" (Kohn, Corrigan, and Donaldson, 2000). This means that not only does the aviation industry have 20 years lead in experience on the medicine industry, but also the nature of the trigger is extremely different: a book vs tragic events. So perhaps the medical industry is simply behind the curve?

In many cases, medical institutions have now implemented systems which attempt to capture and analyse failures in a similar manner to aviation. However, the trust in these systems is not yet present. The prevalence of 'compensation culture' in medicine could be suppressing the reality of implementing a 'just culture'. This is most probably hindered by the application of aviation methods, such as the reporting systems, into the medical sphere without properly accounting for the cultural differences. While there are many similarities to the jobs - high intensity stress, cognitive demand etc - the societal aspects surrounding the industries are not similar.

So, what can construction learn from this?

Systems impact people - but people also impact systems. The comparison of these two industries underline the importance of anticipating and understanding the impact systems have on human behaviour. They also reiterate that systems are affected by the people using them. For example, despite having reporting systems, medical incidents are critically under-reported. Analysis of the resultant data would therefore be biased. It is therefore important to consider the prevalent culture (and the desired one) both when designing systems and when using data they generate.

2.2.4 Should we deal with the root of the failure? Or the leaves? Or both?

Catastrophic failures and collapses, such as those we use for case studies or see in the news, are rarely the consequence of one singular mistake, but rather the accumulation of multiple layers of failure, occurring at multiple levels of the business. Perrow (1999) argued, with engineering examples such as Three Mile Island and Challenger spacecraft, that disasters are rarely the result of one isolated event or system failure but more typically due to a chain or cluster of many interconnected events, either lightly or tightly coupled. Investigation of these interconnected events and how they manifested in failure is also referred to as 'root cause analysis'.

The Swiss Cheese model of accident causation developed by James Reason (Reason, 1990; Reason, 2000), as seen in Figure 2.3, was developed to visualise and analyse the interconnected nature of failure events. In this model, each 'slice' of cheese represents a level of the business which could prevent a failure; meanwhile, the 'holes' are a failure in that level of prevention. For a failure to manifest, there has to be a path through the entire cheese. If any one of those smaller failure events, 'holes', had not occurred in tandem with the others, the larger failure may not have happened. Ideally, it would be possible to isolate these layers and analyse them individually. This would identify the layers which are effective and fail rarely, and which are suffering from systematic failures and require investment to adjust. However, as these layers all fail simultaneously in these case studies, it is impossible to use these cases to determine which layers are failing routinely.

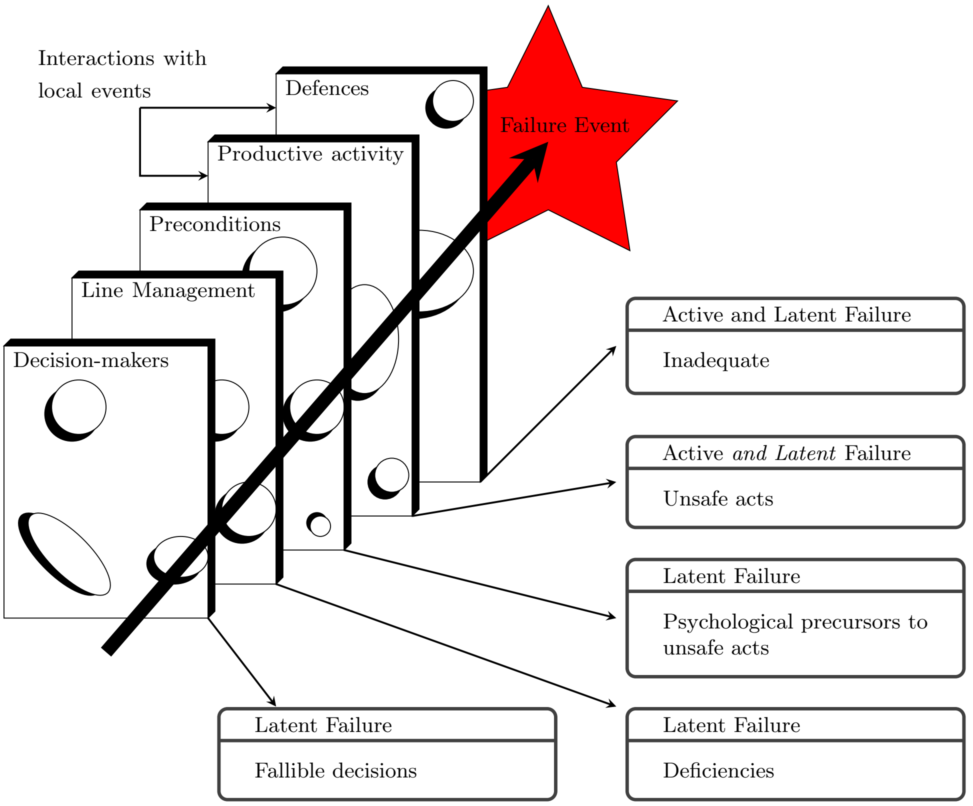


Figure 2.3: Reason's Swiss Cheese Model of Accident Causation (Reason, 1990)

Another model adopted for failure analysis is the Bow Tie method. A comprehensive review of Bow Tie literature by Ruijter and Guldenmund (2015) found that beyond the superficial shape, there was little consensus about the specifics of the method. However, they identify three key preceding methods which combine in some manner. These are: Fault trees, Event trees, and Barrier thinking. Fault trees and event trees both start from a single 'top event', which is generally the point of failure or critical event. While fault trees work back to the causes or hazards which caused that event, event trees work forwards to the consequence. These generally connect to create the left (fault tree) and right (event tree) of the BowTie diagram. Bow Tie diagrams also often incorporate 'Barrier Thinking' into their construction. Ruijter and Guldenmund (2015) state that this makes "an additional distinction between negative events and control mechanisms.. by categorising certain systems or human interventions." Figure 2.4 illustrates a generic Bow Tie diagram.

In both these models, Bow Tie and Swiss Cheese, the manifestation of failure can be pinpointed to a single event. The layers of protection before that point are essential to preventing this event occurring. It could be postulated that the success of each preventative system is a 'failure factor' while the event itself is a 'failure criteria', in similar ilk to the contrast between success factors and criteria discussed earlier.

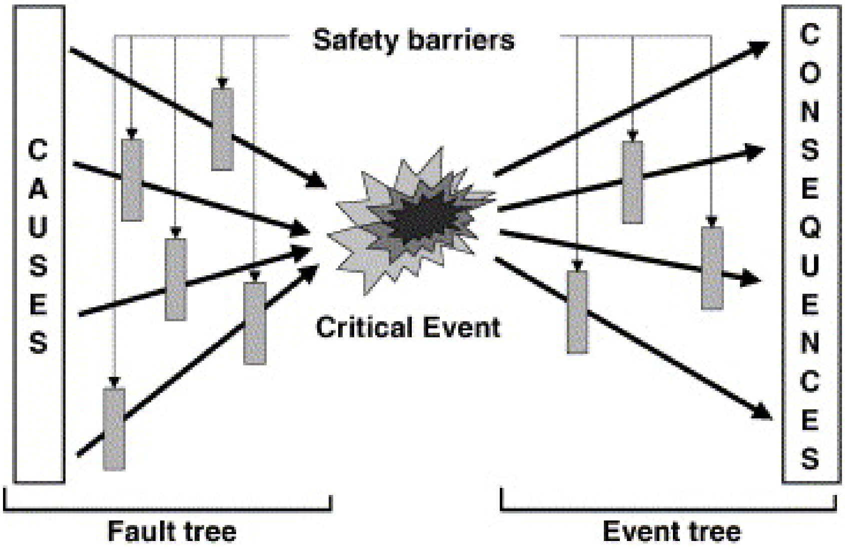


Figure 2.4: Generic Bow Tie (De Dianous and Fiévez, 2006)

An advantage to these methods are that they are extremely explainable - being both straight line logic and graphical. However, as Ruijter and Guldenmund (2015) identified for Bow Tie analysis, there is little consensus in the method of constructing this logic. The events and barriers can be quantitative - for example, using Boolean gates - or qualitative - where descriptive events and barriers are used. Additionally, the selection of the failure event can be difficult. An example of this could be a pollutant spill leads to a fire. Is the spill the top event or the fire?

Nevertheless, root cause analysis has proven an invaluable tool in the analysis of large failure events. An example of this is the investigation into the root causes of the Edinburgh Schools collapse in 2016.

Edinburgh Schools Example

On 29th January 2016, an external wall at Oxgangs Primary School in Edinburgh collapsed. Nine tonnes of masonry fell during a storm onto a pathway used during the school day by children. Luckily no one was injured, but the result could have been fatal.

The independent inquiry into this event concluded that "It is the view of the Inquiry that the primary cause of the collapse of the wall at Oxgangs School was a direct result of poor quality construction, in the building of the external cavity wall which, in the case of a significant proportion of the wall ties failed to achieve the required minimum embedment of 50mm, particularly in the outer leaf of the cavity wall" (Cole, 2017). However, like other cases of catastrophic failure, this singular 'point of failure' was not the end of the story.

Following the wall collapse, surveys were undertaken of 17 schools constructed on the same contract and found a "very significant extent of defective work and omission of components". These schools had been closed as a precaution after the wall collapse and were only reopened after significant remedial work had been completed. The independent inquiry found that "a fundamental weakness of the process adopted was the lack of properly resourced and structured independent scrutiny of the construction and an over-reliance on the part of the City of Edinburgh Council, without adequate evidence, that others in the project structure would comprehensively fulfil this essential role". The manifestation of failure, a wall collapse. had resulted from a combination of organisational choices and shortfalls leading to poor quality compliance.

In the case of the Edinburgh schools, the wall collapse was the manifestation of failure. This prompted the further investigation which revealed failures further up the organisation; the 'holes' in the previous mentioned Swiss cheese. This is quite a unique case in that multiple projects, 17 of which had not exhibited structural failure before intrusive investigation, were investigated and could be used to verify the organisational layer which had failed.

In this case, the benefits of dealing with the 'roots' of the failure are clear and learning from this failure is brought into the arrangements for new projects of this type. However, these investigations rely on the manifestation of a catastrophic failure to warrant the resources invested in them. So, what about if the 'leaves' or manifestation of the failure (the quality issues) had been dealt with as they occurred? Would this have also been a valid method to avoid failure?

Case studies, such as the Edinburgh Schools example, have a track record of extracting lessons to address organisation failure at high levels. In the Swiss Cheese model, these lessons address failure at the decision-makers and line management levels. It appears this kind of activity is appropriate to instigate radical change in a top-down manner. This is discussed further in the next Section 2.3.

However, if failures in the productive activity had been picked up during the construction phase and the activity adjusted, this disaster, and many like it, could have been avoided. While it is insufficient to solely rely on the final layers of defence (Swiss Cheese), it also seems insensible to not also learn to adjust and improve these layers, in conjunction with addressing the root causes. In essence, this project aims to pick out the attributes of smaller failure events to find commonality between these events. This will develop into a learning from failure system for these smaller events - adjustments to the productive activity and defence barriers - and also build momentum for root cause analysis into smaller failures.

2.3 Organisational learning and learning from failure

2.3.1 What organisational learning theory is most relevant?

Organisational learning research covers a wide range of disciplines and research methodologies. Easterby-Smith, Crossan, and Nicolini (2000) acknowledge that there has been much past debate about what ‘organisational learning’ actually entails, mostly revolving around whether ’learning’ denotes cognitive or behavioural change. However, it is now generally accepted that it could be either or both, and this remains a point of definition and departure for organisational research. With this in mind, this section explores past and present themes.

Past

Argote (2011) identifies Levitt and March (1988)’s “Organizational Learning” article in the *Annual Review of Sociology* as a breaking point between past research and themes from 1980s to 2010. She noted that before the 1980s there were three, mostly independent, fields of inquiry and provided examples of each. These fields were:

1. Learning curves - research mainly performed by engineers and economists concerned with how performance characteristics changed with experience.
2. Human-centred barriers - predominantly psychological case studies and clinical research exploring how human defensive routine prevent learning.
3. Theory development for learning as changes in organisation’s routines which influence behaviours - a sociological field based mainly in simulation work, originating from Cyert and Marsh (1963).

After this point, Argote (2011) goes on to identify three dominant themes which emerge between 1980 and 2010. These are: experience, context and processes. To a certain extent, experience and processes are a development of the previous fields of learning curves and organisational routines respectively. Exploring examples of work in these areas, a distinct departure from previous work is the inclusion of other methodologies and methods. Argote (2011) indicates that this is from some co-mingling of the original disciplines, however, ‘context’ does not directly originate from a blend of these prior fields but rather from a “more socially aware stance between learning and knowing” (Easterby-Smith, Crossan, and Nicolini, 2000).

Present

Since 2010, the three themes identified have developed. In particular, four key dimensions to these fields have been identified in the literature which are particularly relevant for this research. These are discussed here and include:

- Context
  - Organisational learning climate
  - Levels of learning
- Experience
  - Types of knowledge
- Processes
  - Digital learning

Organisational learning climate

Within consideration of the *context* theme, a key debate surrounds the existence of organisational learning culture or climate. Culture and climate are deeply entangled concepts which have sparked complex ontological debate over the years. In recent literature, organisational climate is widely accepted as a manifestation of organisational culture, where organisational culture refers to an intangible set of underpinning values in a company, influencing its decisions and business architecture, while organisational climate refers to the employees' perception of the values, processes and priorities of the business. In other words, organisational climate is a visible manifestation of the culture (Ekvall, 1996).

Thus, learning organization climate is the employees' perception of the values, processes and priority of organisational learning in the company. It should be noted that *organisational learning* and *learning organization* are still used almost interchangeably despite observation and efforts to separate the two, for example by descriptive vs prescriptive meanings (Anders Örtenblad, 2001; Easterby-Smith, Crossan, and Nicolini, 2000). In other words, *learning organization* describes an entity which possesses characteristics which systematically support change in response to events and technological changes. Meanwhile, *organisational learning* is the process of collective change of cognition or behavior within the organization (Argote, 2011). While *organisational learning* can refer to negative change, for example corporate forgetting, it is generally referred to in literature as the positive version of the behavior exhibited by organisations.

Important characteristics, known as dimensions, which learning organisations are said to exhibit have been collated in previous literature. These are key to understanding what is important when designing a systems for learning. Here, two sets of dimensions for learning organisation climates are presented in Table 2.3. The first, the 'Dimensions of Learning Organization Questionnaire' (DLOQ) by Marsick and Watkins (2003), has been used in different industries around the globe, both in academic studies and practice. The second is that of Garvin, Edmondson, and Gino (2008), who developed a survey which was aimed more for assessment in commercial organisations.

One of the dimensions identified is to possess technological processes which collect and analyse information to discover lessons - as is aspired to in this research. However, it is key to also consider the other dimensions in the creation and promotion of new learning systems. For example, it should be considered whether new systems encourage dialogue or dictate rules; whether they empower people or micromanage them; whether they connect the organisation or encourage insular behaviour.

Additionally, there are certain limitations to identifying important dimensions for organisational learning using the epistemological view adopted by this approach. The most significant is that dimensions independent of the employees' perception (for example unknown dimensions, other leading or lagging indicators) may be excluded from the analysis. In other words, any learning climate measure obtained via staff questionnaires may lean towards the staff development aspects, neglecting high-level learning implementation of policy or process revision. Therefore, these aspects should be considered alongside the possibility of other indicators of learning, such as evidence of processes and performance indicators.

Levels of learning

Easterby-Smith, Crossan, and Nicolini (2000) noted that early researchers indicated that organisational learning could be described by mapping the learning and cognitive behaviors of employees, or senior managers. For example, Fiol and Lyles (1985) suggested that organisational learning climate describes the factors affecting individual learning. However, this would imply that the organization is a sum of its employees. Other research has disputed this, acknowledging that while individual learning is part of the picture, there are other mechanisms at play. As phrased

Table 2.3: Dimensions of a learning organisation

Dimension		Description*
Marsick and Watkins (2003)	Garvin, Edmondson, and Gino (2008)	
Create continuous learning opportunities	Education and training	Formal processes exist for learning on the job; opportunities are provided for ongoing education and developing employees' skills.
Promote inquiry and dialogue	Psychological Safety Appreciation of Differences	People express their views and listen and inquire into the views of others. Employees feel safe disagreeing with others, asking naive questions, and owning up to mistakes.
Encourage collaboration and team learning		Work is designed to use groups to access different modes of thinking; collaboration is valued by the culture and rewarded.
Create systems to capture and share learning	Information collection Analysis Information transfer	The organisation has both high- and low-technology processes to generate, collect, interpret, and disseminate information; access is provided; systems are maintained.
Empower people toward a collective vision		People are involved in setting, owning, and implementing a joint vision; responsibility is distributed close to decision making so that people are motivated to learn toward what they are held accountable to do.
Connect the organisation to its environment		People are helped to see the effect of their work on the entire enterprise; people scan the environment and use information to adjust work practices; the organisation is linked to its communities.
Provide strategic leadership for learning	Leadership that reinforces learning	Leaders demonstrate willingness to engage in active dialogue and entertain alternative viewpoints; signal the importance of spending time on learning; and use learning strategically for business results.
Experimentation	Openness to new ideas	Employees take risks and explore the unknown. The organisation has processes for experimenting with new offerings.
Time for reflection		Employees take time to review organisational processes.

*Descriptor based upon combination and summary of two reference descriptors.

by Argote (2011), "the individual's knowledge would have to be embedded in the organization so that other members could access it, even if the individual left the organization".

It is therefore important to make the distinction between facilitating individual learning at the workplace and organisational learning when considering the dimensions collected. There are now four acknowledged levels of learning within an organisation: personal, group, organisation and inter-organisation. These levels require different approaches and systems should be adapted accordingly.

Not only is it important to tailor learning systems for each level of learning, it is also important to appreciate the interaction between these levels. A recent qualitative study by Morland, Breslin, and Stevenson (2019) explored the interaction of these learning levels for a large UK housebuilding company. They found that communication between these levels synchronised over time, leading to better sense-making. Communication between these levels is referred to as *feed-forward* (up the management hierarchy) and *feedback* (down the management hierarchy). Trust was key in facilitating this communication so any learning systems should ensure that it builds, rather than undermines, trust.

Types of knowledge: tacit vs explicit

A foundation of learning theory is the clarification between different forms of knowledge. A key distinction is that between tacit and explicit knowledge. Choo (2000) defines these where explicit knowledge is comprised of facts and information which are either rule-based (e.g. in an organisation's rule, regulations and procedures) or object-based (e.g. in products or data stored by the organisation). Meanwhile, tacit knowledge is the personal knowledge 'stored' within the organisation's personnel. An example given by Choo (2000) is "the bank manager who gets a gut feel that a client would be a bad credit risk after a short conversation with the customer". They state that "since tacit knowledge is experiential and contextualised, it cannot be easily be codified, written down or reduced to rules or recipes."

These different types of knowledge are integrated into organisational learning in different ways. While it may appear from these descriptions that only explicit knowledge is suitable for organisational learning as it can be captured and communicated via written documents, tacit knowledge can (and is) also integral to the function of many industries, including construction. In considering organisational learning systems, it is essential to define the type(s) of knowledge being dealt with. Specific ways in which different knowledge is captured and stored by organisations is covered during the next section in consideration of knowledge management systems.

Digital learning

For literature from the 90s to early 00s, information technologies such as personal computers had just become the norm in office space. Now, in the UK, 60% of employees use computers with internet access for work. This is radically changing the systems used for organisational learning and has anchored the realm of organisational learning to that of ICT (information and communication technology) and management of the organisation's data (part of its explicit knowledge base).

Online or distance learning is on the rise. Equally, company policy can be changed and disseminated across all levels of business cheaply and easily. Digital technology affects all levels of organisational learning and for all knowledge types. New systems will have to be adaptable to incorporate technological advances and should exploit the opportunities afforded with the accessibility of this technology. To this end, this research actively engaged with what technology is used in construction projects and aimed to exploit the rise in access to both digital data and connect it to people.

2.3.2 What is learning from failure?

What do we mean by learning from failure? As infants, we learn instinctively from past experiences. Learning to avoid certain behaviours and increase others to achieve the desired effect, for example, learning to walk where one may fall many times before learning how to walk successfully. Kolb (2015) defines this type of learning, where lessons are extracted from the ordinary course of life, as 'experiential learning'.

However, converting this learning to an organisation or industry is notoriously difficult and significant learning was historically limited to large public failures. This difficulty can be attributed to a combination of the technical complexity with implementing continuous learning in an organisational context coupled with negative social and psychological reactions which most people exhibit when faced with the reality of failure (Cannon and Edmondson, 2005). Evidence of the difficulties of learning from failure, even large failures, can be observed in the recurrence of similar failure types.

This subsection explores relevant organisational learning theory related to learning from failure and learning from failure in practice where failure can be described as undesirable or unintended outcomes.

A seminal theory for learning from failure is that of single and double loop learning, first introduced by Argyris (1977). This is illustrated in Figure 2.5. The single-loop learning cycle focuses on correction of a procedure or behaviour to prevent recurrence of the failure mode but does not examine the underlying values. Argyris (1977) uses the example of adjusting the temperature instruction given to a thermostat to correct the failing of a cold room. The instruction is corrected to prevent failure; however, the values and culture behind the process are not questioned, e.g. they did not ask if donning a jacket (something outside of the system norms) would achieve the same goal more efficiently. If this extra loop is included, Argyris (1977) refer to this as double-loop learning. It is important to appreciate that this model was first developed as a 'Theory of Action' to explain the rationale behind human decision making. By examining 'inhibiting loops', it is used to identify and explore the consequences of barriers to learning. However, since then, the skeleton of this model has been applied and proven relevant for experiential learning in many contexts and levels (individual, group, organisation, inter-organisation). For example, Drupsteen and Hasle (2014) aggregated information about learning from failure in organisations to create a generic stepwise learning cycle which was identical in basic form to that of Argyris (1977).

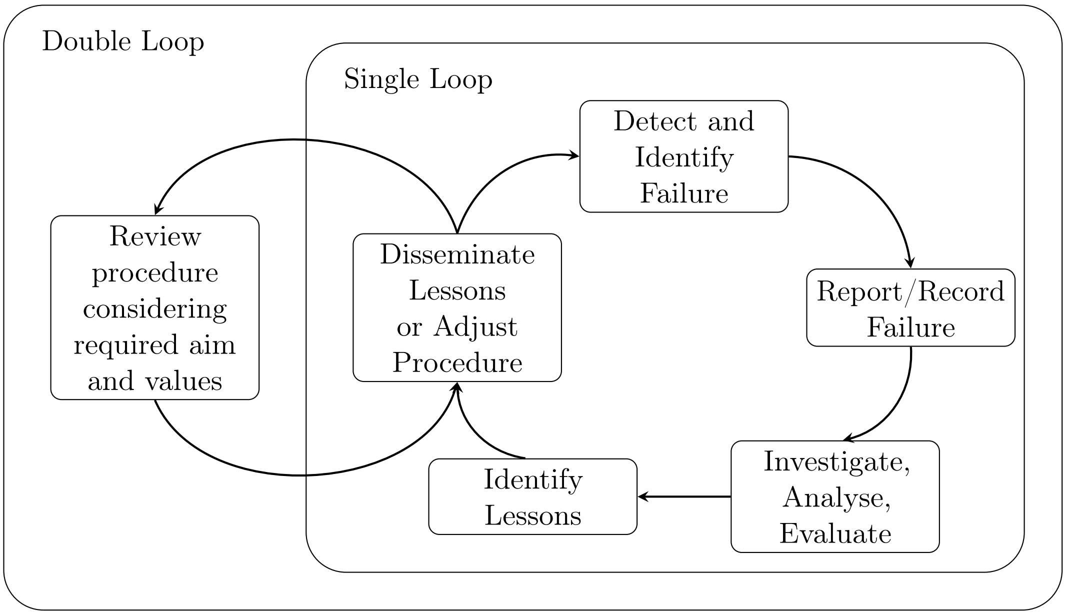

Double Loop
Single Loop
Detect and Identify Failure
Report/Record Failure
Investigate, Analyse, Evaluate
Identify Lessons
Disseminate Lessons or Adjust Procedure
Review procedure considering required aim and values

Figure 2.5: Single vs Double Loop Learning (Argyris, 1977)

In comparing processes which exhibit double or single loop learning characteristics, double-loop learning is often referred to as superior; with Stemn et al. (2018) suggesting that classification of whether an implemented learning system included and/or encouraged double-loop learning could help define the effectiveness and maturity of the cycle. There is a certain irony to this, however, as if a learning system *systematically* includes a double-loop like review into the procedure, this review becomes part of the single loop process. Therefore, it could be postulated that a formal process, on its own, can never exhibit double-loop learning.

Easterby-Smith, Crossan, and Nicolini (2000) note that there is criticism that the classification of single and double loop learning for organisations is paradoxical as some believe that "double loop learning requires *outside* intervention to make it work, and yet judgments for the need for double loop learning can only be formed from *inside* the organisation which is, by definition, locked into a process of single loop learning". While this is a valid concern, it ignores the context of organisations. Organisations do not operate in a vacuum and the people working within them are not solely cogs in the organisation. If they were, it would probably be true to say that organisations are "locked into single loop learning" as there would be no external thought process to realise that a change in norms would be beneficial. However, people are also cogs in their own personal lives and in society. They are therefore exposed to values and cultural norms external to the organisational setting, and this exposure helps form judgments.

In conclusion, the single/double loop learning model is a relevant model to facilitate systematic exploration of processes and barriers to learning from failure in organisations. Subsequent discussion is broken into aspects of this and identification of barriers to learning.

Recently, much research on 'learning from failure' has revolved around investigation of the concept in practice, mainly based on case-studies. Recent reviews, such as Drupsteen and Hasle (2014) and Stemn et al. (2018), have shown limited implementation of learning from failure within industry, and research has focused on identification of barriers to this learning. This research has been based both in general organisations or engineering projects, for example (Cannon and Edmondson, 2005; Drupsteen and Hasle, 2014; Stemn et al., 2018). These investigations have used models similar to, or identical to, the learning loop model from Argyris (1977) to identify these barriers.

Stemn et al. (2018) identify, from 40 peer reviewed papers, the following barriers in four categories to learning from incidents:

- Learning inputs:
  - Non-detection and non-identification of reportable incidents
  - Under-reporting of detected incidents
  - Lack of focus on small precursor incidents
- Learning process:
  - Inadequate description of reported incidents
  - Superficial investigation and analyses of incidents
  - Poor selection, planning and implementation of corrective actions
  - Lack of effective learning from incidents (LFI) systems and sharing lessons
- Learning context:
  - Culture of blame, lack of trust and expected performance created by management
- Learning agents
  - Beliefs, experiences and competencies of actors of learning

Specific to learning process, a regularly reported barrier to learning from failure is the lack of accessible information on past failures, for example (Cannon and Edmondson, 2005; Drupsteen and Hasle, 2014; Stemn et al., 2018). This is especially relevant for construction failures considering the volume, variety, fragmentation and confidentiality of the information involved. This barrier is returned to in Section 2.4.

This learning relies on individuals identifying what they believe to be significant cases of failure on their project, either for their general applicability or potential consequences, and then disseminating this information to a wider audience. Communication of this failure often takes the form of an alert or storytelling, either to an individual via IT or by forums. Silva et al. (2017) also identified two further intervention strategies used to implement learning, in addition to diffusion and discussion highlighted above. Training refers to the use of incident information to improve or introduce employees' training, while change describes the adjustment of a procedure or standard in response to an incident. These are both top-down approaches instigated by leadership.

Additionally, while identification of barriers to learning from failure has taken place in wider context, specific examination of the construction setting is lacking, especially in regards to how established processes interact with the attitudes of employees. The next sub-section (Section 2.3.3) examines the literature specific to construction which exists.

A side note on innovation management

A term emerging in popularity attached to experiential learning is 'innovation management'. This type of learning aims to develop new ideas or processes. Innovation "..is an idea, practice or object that is perceived as new by an individual or the unit adopting it." (Rogers, 2003 p.16). Closely related to 'agile business' and 'experimentation', this term has gained popularity after the success of Silicon Valley start-ups and giants such as Apple and Google. At its core, it is a mode of experiential learning which is often quoted to aim to "fail fast and fail often".

It is necessary to differentiate this from 'learning from failure' investigated in this research. There is an underlying philosophical difference in the aims between 'innovation management' and what is referred to here as 'learning from failure'. In the first, the aim is to learn how to do something specific, however, the second aims to learn how to avoid a failure event in existing processes or tasks. For example, if we were walking, we would be trying to avoid falling down rather than learning to walk. This leads to an interesting philosophical debate about the underlying methodology of learning in the construction industry. This debate is picked up later in the discussion and limitations section.

2.3.3 How does learning from failure manifest in construction?

As will be seen, there is little literature available concerning organisational learning in construction. There are even fewer pieces exploring learning from failure. It was identified early into this research that, in order to properly inform the investigation, more information would be required than could be gleaned from the existing studies. Therefore, the literature presented here is supplemented with a preliminary investigation, presented in Chapter 4. After the start of this research, Lundberg, Lidelöw, and Engström (2017) also reached this conclusion during their systematic literature review, which worked from the question "What are the methods for organisational learning in terms of knowledge sharing and knowledge transfer in the everyday practice of construction projects on site level in a western world context?". They also surmised that there is a need for further study of both current practice and context, as well as development of theory.

So, what do we know? As with most industries, inter-organisational learning in construction is historically from large public failures, such as that of the Tacoma Narrows Bridge or Hyatt Regency Hotel walkway. The root cause investigations which follow such tragic or infamous events are extensive. As presented previously (see Section 2.2.4), these events are rarely caused by a singular mistake and are often the result of a combination of several failures, at different levels of the business. The tacit knowledge gained from these investigations is then incorporated into standards and guidelines. The implication is that this event will not happen again as the rules now prevent it. These events are, thankfully, relatively infrequent. Therefore, this learning process (whereby failure informs change) is equally infrequent and informed by few data points.

Another method of learning from failure in construction is by studying failure case-studies. Curated databases exist of case study examples, such as those included in the SCOSS (Standing Committee for Structural Safety) database or anthology publications (Breysse, 2012; Soane, 2016). These contain cases considered significant for development of personal knowledge and industry learning. They represent a high detail, low volume dataset which is accessed on a case-by-case basis for learning. These characteristics also apply to lessons learnt databases in individual organisations. Lessons learnt documents are considered an essential part of the knowledge management system which aim to add value to the organisation by capturing and disseminating lessons (Caldas et al., 2009). In both these sets of data, curated case-studies and lessons learnt, there is an reliance that individuals will engage with the process and search for previous lessons when undertaking a new task. Lampel et al. (2009) dub this type of engagement 'learning about failure' rather than 'learning from failure', which highlights a key distinction in the level of engagement involved.

Along the theme of 'learning about failure', recent research has explored the importance of storytelling on construction projects, especially for implicit knowledge. In a five month ethnographic study of railway construction, Sanne (2008) found that storytelling was "*an integral part of technicians' practices and their accident etiology and creates a way for them to address risks*", while the formal incident reporting process was neglected due to lack of everyday relevance. However, Sanne (2008) also notes that the reliance of anecdotal stories only addresses these failures from a narrow perspective and neglects root causes. [1]

From this, it appears that the construction industry places learning as the responsibility of the individual. While this has been proven to be inefficient in other industries, perhaps the construction industry facilitates this in such a way as to make it work. Previously presented is the concept of learning climates, a set of conditions which actively encourage learning in the workplace. So, is there evidence of this in construction? Regretably, not. This research found only one example of a construction specific organisational learning dimension survey by Kululanga, Price, and McCaffer (2002).

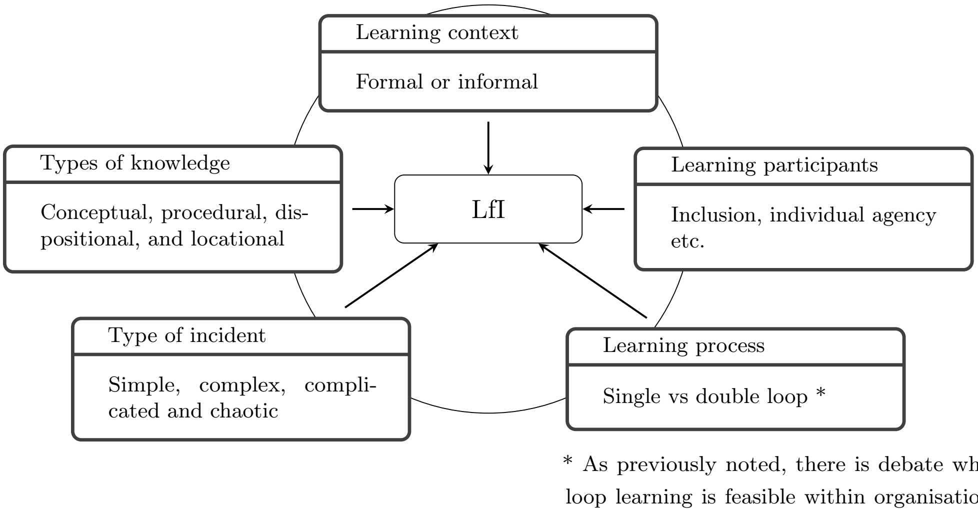

Learning context
Formal or informal
Types of knowledge
Conceptual, procedural, dispositional, and locational
LfI
Learning participants
Inclusion, individual agency etc.
Type of incident
Simple, complex, complicated and chaotic
Learning process
Single vs double loop *
* As previously noted, there is debate whether double loop learning is feasible within organisations.

Figure 2.6: Learning from Incidents (LfI) Framework (Lukic, Littlejohn, and Margaryan, 2012)

Perhaps the most relevant field of research concerning learning from failure is that of Learning from Incidents (LfI). While not based within the construction domain, it focuses on learning from safety incidents in high hazard industries, such as construction. A series of papers published by Littlejohn, Lukic and Margaryan explore and develop recommendations for LfI processes, e.g. Littlejohn et al. (2017), Margaryan, Littlejohn, and Stanton (2017), and Lukic, Margaryan, and Littlejohn (2013). Their framework is presented in Figure 2.6. As seen, understanding of learning depends on factors inherent to the specific industry or organisation, such as participants, knowledge type and learning context. Therefore, while this research can be used as a starting

[1] While this study is limited in generalisability due to its methodology, the prevalence in storytelling for learning on site is also reinforced by my own experience and is further investigated later in this thesis.

framework, it is essential to validate these theories in the construction domain. Systematic reviews of LfI literature, such as Drupsteen and Guldenmund (2014), also concur with this assessment.

In conclusion, the limited literature available seems to imply that learning in construction focuses on building personnel competence and change is driven by the people themselves. Yet, an apparent lack of consideration of learning climates suggests that this is not due to deliberate thought but rather by neglect of other processes. Meanwhile, organisational learning occurs through adjustment of standards. In relation to failures, this means that construction learns from catastrophic events or case studies, and relies on individuals' intervention to attempt to incorporate and communicate learning from smaller events - either through alerts or stories. There appears to be a lack of formal processes connecting the communication of failures to implementation of change. There also appears to be a lack of methods dealing with high volumes of low detail data. This is discussed further in Section 2.4.

Additionally, these conclusions are drawn from the lack of information or literature rather than literature itself. For this research to be most effective, it first needed to be well informed about the current systems in place. Therefore, the lack of literature on this topic necessitated a preliminary investigation, presented in Chapter 4.

2.4 Knowledge Management (KM)

Knowledge management (KM) lies within the intersection of organisational learning and information technology. In the last 30 years, this has become a field of research in its own right and is considered crucial to the success of an organisation (Asrar-Ul-Haq and Anwar, 2016). This section gives a brief introduction to knowledge management and its connection to organisational learning before exploring key KM theory themes in relation to learning from failure. This is followed with a critical discussion of knowledge management in construction.

2.4.1 How does knowledge management link to organisational learning?

Knowledge management research and application has grown significantly since the mid-90s. Akhavan et al. (2016) found in their bibliometric study of KM literature that published article numbers grew from <50 papers per year before 2000 to over 350 in 2007 where the growth plateaued. During this period of growth, researchers have looked to define different types of knowledge, and the value it has. A well-established hierarchy of knowledge by Tuomi (1999) is shown in Figure 2.7. This relates the knowledge level to the ability to make decisions or take actions from it. KM processes deal with knowledge at all these levels and the transfer between them - not just at the 'knowledge' level. This could lead to some confusion during discussion of processes and content - unless otherwise stated, in this section, 'knowledge' refers to any level of the hierarchy.

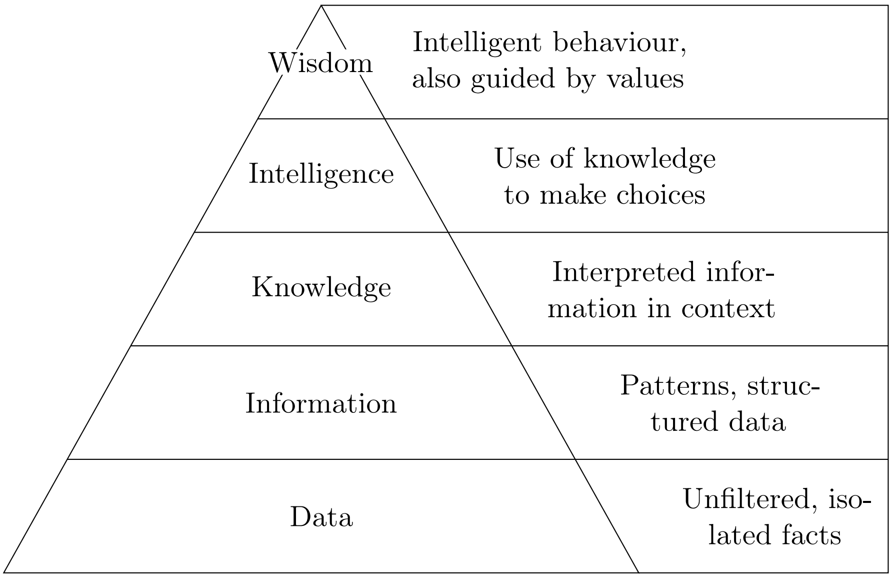

Wisdom
Intelligent behaviour, also guided by values
Intelligence
Use of knowledge to make choices
Knowledge
Interpreted information in context
Information
Patterns, structured data
Data
Unfiltered, isolated facts

Figure 2.7: Hierarchy of Knowledge (Tuomi, 1999)

According to Ouriques et al. (2018), the main processes of KM are: knowledge creation, knowledge storage/retrieval, knowledge transfer/sharing and knowledge application. It is not a coincidence that these steps are similar to those of organisational learning (see Section 2.3.2). In fact, the learning cycle can be seen as the process of moving up this hierarchy to facilitate intelligent (single-loop) or wise (double-loop) action as seen in Figure 2.8. By combining these models, the relationship between the concepts of organisational learning and knowledge management is illustrated clearly: knowledge management processes facilitate movement around this learning cycle.

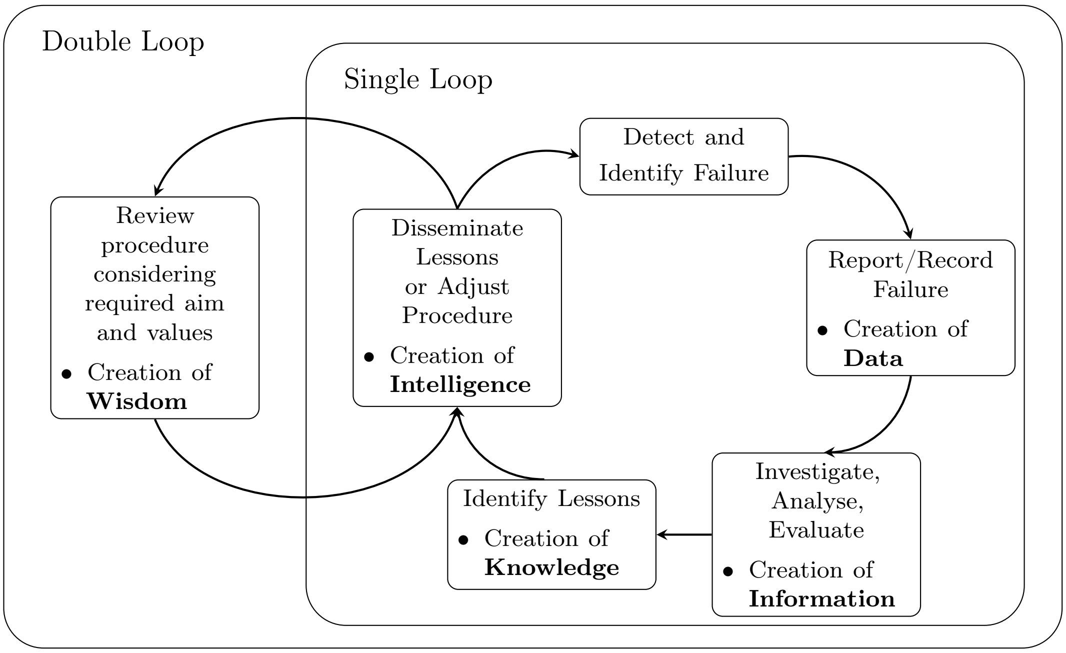

Double Loop
Single Loop
Detect and Identify Failure
Report/Record Failure
• Creation of **Data**
Investigate, Analyse, Evaluate
• Creation of **Information**
Identify Lessons
• Creation of **Knowledge**
Disseminate Lessons or Adjust Procedure
• Creation of **Intelligence**
Review procedure considering required aim and values
• Creation of **Wisdom**

Figure 2.8: Learning Cycle (Argyris, 1977) in relation to Hierarchy of Knowledge (Tuomi, 1999)

This research is not the first to draw parallels between organisational learning and knowledge management research. Easterby-Smith and Lyles (2011) initially present the distinction between knowledge and learning as a scale of content to process, as shown in Figure 2.9. They describe this as "knowledge being the stuff (or content) that the organization possesses, and learning being the process whereby it acquires this stuff", however, they soon point out the inadequacies of this binary classification method and clarify that they present such an oversimplified view to have initial organising principles with which to work from. This is in line with the philosophy already presented for such management tools - these are intended to prompt discussion and organise thoughts rather than as a prescriptive framework. Indeed, this relationship does not reflect the nuances of either field. By presenting these concepts as a dichotomy, they undermine the symbiotic nature of the two fields where both fields benefit from application and development of the other.

A different review article, nearly a decade later, found that this relationship has further evolved. In their comprehensive review of over 16,000 articles, Castaneda, Manrique, and Cuellar (2018) conclude that organisational learning is being conceptually absorbed by knowledge management research. They note that during the 2006-2014 both fields of research displayed a increased interest in "linking learning and knowledge with organizational strategy, results and competitiveness" as well as "in understanding the role of organizational culture". They found that the relationship between the key terms and theories show that, rather than being sub-concepts of a larger research field, organisational learning is becoming a sub-concept of knowledge management.

Over the course of this research, it became clear that knowledge management was inextricably linked to the concept of organisational learning and was essential for this research. Additionally, knowledge management also deals with other applications of knowledge in organisations, such as organisational strategies for document retention and access for tasks like quality assurance and evidence if future problems arise. These tasks are extremely relevant for failure data and affect the data collected - both in terms of ideal content and human factors which affect the content/biases of the data. Therefore, understanding how KM relates to learning from failure is

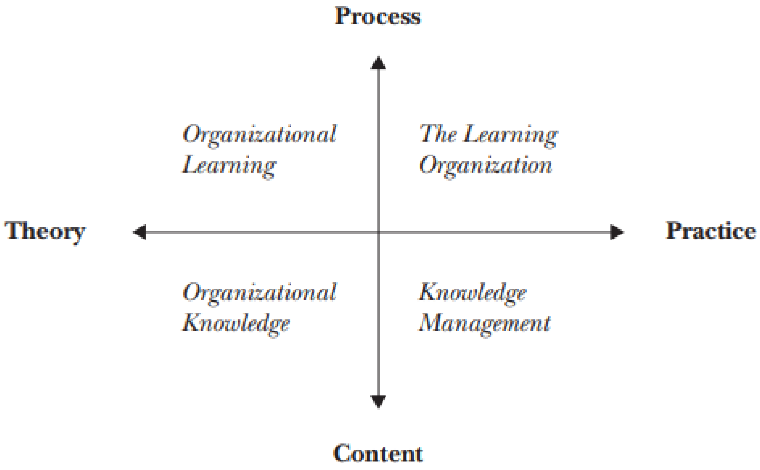


Figure 2.9: Organisational field and knowledge management (Easterby-Smith and Lyles, 2011)

essential for this research.

This research adopts Inkinen (2016)'s definition of KM as "the conscious organizational and managerial practices intended to achieve organizational goals through efficient and effective management of the firm's knowledge resources". This intentionally excludes 'organic' forms of knowledge management - that is knowledge management processes which develop without deliberate thought or strategic aims.

2.4.2 How does knowledge management apply to learning from failure?

As established in Section 2.3.3, failure events can instigate learning via a single-loop or double-loop process; however, organisational learning research has identified multiple barriers to realising positive action(s) from these events. Knowledge management principles aim to address these barriers, from the angle of managing the knowledge resource via technology and human resources.

Do failure events create explicit or tacit knowledge?

The short answer here is 'both'; however, understanding the nature of the knowledge or data produced by failure is essential to understanding the KM process. This is because a key definition dictating the knowledge management methods appropriate for facilitating learning is the type of knowledge being handled. Heisig (2009) found that 35% of KM frameworks in his review dealt directly with the dichotomy of tacit and explicit knowledge.

In a seminal piece of KM research, Nonaka (1991) wrote that "new knowledge always begins with the individual". His argument was that knowledge always began as tacit knowledge gained from an individual's experience which can then be transferred. Instead of focusing on the 'elevation' of knowledge up a hierarchy, his article introduced 'The Spiral of Knowledge' - a set of four processes which describe the transfer of tacit and explicit knowledge. The four processes were:

1. Socialisation: tacit to tacit. Whereby others can gain knowledge by observation, discussion and apprenticeship.
2. Articulation or externalise: tacit to explicit. Nonaka (1991) notes that this can be capturing the knowledge as instructions or data external to the person i.e. in a document or by creation of a product/procedure/standard which captures this knowledge.

3. Combining: explicit to explicit. By combining explicit data and synthesising this combined data sets, Nonaka (1991) said an individual can perform explicit to explicit knowledge transfer but that this does not create new knowledge. This is the least defined step and is discussed further below.
4. Internalise: explicit to tacit. Where personnel receive an explicit piece of information and use it to "broaden, extend and reframe their own tacit knowledge".

There are two main reasons why these concepts are unsuitable for direct use here and need critical re-framing to bring value to this research. The first is that Nonaka's framework was developed with the aim of innovation in mind. To return to an earlier point, learning from failure is not innovation management; the two are philosophically different concepts. The second is that modern digital technology has radically changed the knowledge management landscape since 1991. Therefore, the value in Nonaka (1991)'s four process types lies in appreciation of directional knowledge transfer between these two types - tacit and explicit knowledge.

Returning to the creation of new knowledge, Nonaka (1991) specifically stated that this can only be achieved by human-beings. He described eureka-type moments which lead to innovation and new products/processes. The type of knowledge he described is contextualised and interprets the information surrounding a topic; it is 'knowledge' insofar as Tuomi (1999) defines it in his hierarchy. As seen in Figure 2.8, this is not the start of the learning from failure cycle. In fact, Nonaka (1991)'s own vignette describes several steps before 'knowledge creation' where a failure is identified and reported - before the innovation process begins. This could suggest that innovation frameworks and strategies are a sub-process of learning from failure - i.e innovation is taking the lessons/knowledge and applying it to adjust or review then reinvent procedures. Again, this debate is returned to in the discussion and limitations section.

In this case, using proof-by-example, modern sensing equipment and automated methods can identify and record failures without human intervention, creating explicit failure data; meanwhile, human experience of a failure creates tacit knowledge enriched by context and the individual's previous experiences. Articulation of this knowledge (tacit-to-explicit knowledge transfer) can then be employed to report or record this tacit knowledge and it becomes explicit data - isolated facts - nb. there is always information lost on these interfaces. Therefore, both explicit data and tacit knowledge can be generated by failure. [2]

This tacit knowledge can be directly applied by those involved - this is personal experiential learning - however, explicit data requires further steps to make useful. Company strategies often focus on the management of this explicit data.

How do KM strategies facilitate learning from failure?

Hansen, Nohria, and Tierney (1999) defined two aspects of KM as the codification and personalisation processes. These are strategies of KM focusing on human-technology and human-human interfaces respectively. Their research emphasises the role of people - humans - in the knowledge management process, even when technology is a facilitator. Human involvement remains a key factor in KM literature - Heisig (2009) found that human-orientated factors were the most mentioned critical factors, mentioned in 100/119 KM frameworks, while the individual-collective knowledge dichotomy was the second most frequently discussed dichotomy (behind tacit-explicit knowledge).

[2] So, maybe Nonaka was correct in his statement that knowledge (as defined in the hierarchy) can only be generated by human as the process of 'interpretation' is human-based. While this is based on a single example and not conclusive, AI processes (as explained in the next section) are limited to using patterns in specific data sets and lack general understanding required to interpret what this means - they can generate insights but not interpret them. AGI (artificial general intelligence), which would achieve this, is a sci-fi pipe-dream.

Gammel et al. (2019) argue that focus on the interplay between technology and human-factors has led to focus on socio-technical systems (STS) methodology for KM processes, such as presented in their simplified framework in Figure 2.10. Considering knowledge management processes through an STS lens deliberately emphasises both physical and social outcomes. Design of KM systems using this philosophy should aim for 'joint optimisation' so that both parts yield positive outcomes ("Socio-technical systems theory: an intervention strategy for organizational development").

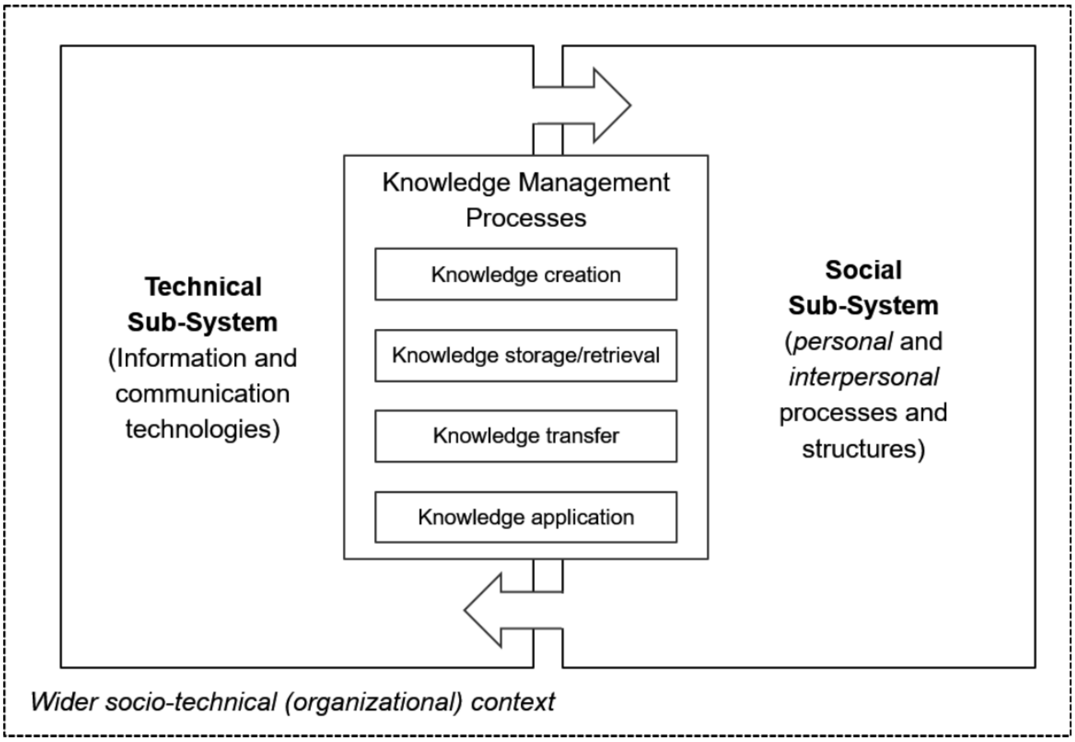

Knowledge Management Processes
Knowledge creation
Knowledge storage/retrieval
Knowledge transfer
Knowledge application
Technical Sub-System (Information and communication technologies)
Social Sub-System (*personal* and *interpersonal* processes and structures)
Wider socio-technical (organizational) context

Figure 2.10: Simplified Socio-Technical Framework for KM (Gammel et al., 2019)

KM aims "to achieve organizational goals through efficient and effective management of the firm's knowledge resources" (Inkinen, 2016). As already established, this knowledge resource includes explicit knowledge stored in documents and artefacts, as well as knowledge within the human resource. While the technology has rapidly evolved since the millennium, the two strategies focusing on human-technology and human-human interfaces still hold true. Organisations make a conscious decision to invest in KM technology or to emphasise human-human knowledge transfer - both require time and money to implement systematically and effectively.

KM strategies focusing on codification have tended to invest in and develop technological solutions to develop databases of relevant information and facilitate suitable access methods. This could be a 'library' of reports, but could also be process flowcharts and other artefacts which capture explicit knowledge. Companies should consider the aims of such systems, the technology available and which technology best fits with their people and organisation.

A UK example of failure to implement such a system is the NHS's abandoned 'National Programme for IT', a £6billion project of which a large part was an electronic patient record system. This system aimed to centralise storage and access to patient records. Currie (2012) concluded that this focal part of the scheme was critical in the failure of the entire programme. However, she stated the technical system "played a relatively insignificant part as the story unfolded" and that the failure of this project was primarily due to human-factors - including lack of stakeholder involvement in the solution design and buy-in for the decisions. This case study demonstrates the importance of considering both the technical and social aspects of KM systems.

In considering single-loop learning (see Figure 2.8), these library-type KM systems effectively by-pass the bottom half of the learning cycle - simply storing and disseminating documents. KM systems which systematically analyse data and extract lessons/insights are less frequent, especially considering unstructured/text data. However, with the rise of automation, modern informatics and data analysis methods, these tasks are becoming more systematically included within KM systems. This is discussed in the next subsection.

Discussion has focused on human-technology systems, however, personalisation remains an important aspect of organisational KM. Hansen, Nohria, and Tierney (1999) uses the examples of consultancy services which focus on dialogue and knowledge transfer via workshops and one-to-ones rather than digital solutions. When these methods are built systematically into organisation culture and policy, they become part of the KM strategy. However, sometimes reliance on these human-human methods evolves without deliberate thought or is significant only due to lack of other technological support. In this case, returning to the definition of KM, these processes are not part of a KM processes.

It should also be appreciated that every interface, be it human-human or human-technology (or technology-technology as in automated processes), loses and gains information. Abstraction processes - where rich experiences or events are 'abstracted' into salient features or numerical data - lose the contextual richness but gain the ability for combination and analysis. Meanwhile, 'contextualisation' processes interpret data in context to form lessons and motivate action. This gains insight, but loses granularity and nuance.

How has modern technology transformed knowledge management?

Many KM processes have been changed beyond recognition in the last 30 years by rapid development of ICT (Information and Communication Technology) and personal computers. When Hansen, Nohria, and Tierney (1999) wrote about codification processes, they describe knowledge being "carefully codified and stored in databases, where it can be accessed and used easily by anyone in the company". At the time, this basic task had already been radically changed by the increased ability and affordability of computers. Prior to computer databases, this knowledge would be captured in typed or written reports and filed in some form of records room - accessed physically and searched using complicated filing systems.

The vignette below accounts my dad's experience of digital technology development in the workplace. This illustrates the changes brought by ICT processes from the 70s to the early 00s. This section will then go on to explore modern models of ICT for knowledge management.

A family history of computers - 'what's a window?'

Gary Winsor, my grandpa, said of computers that "every three years, they will be three times faster, a third the size and a third the cost". He said this to my dad in the 1970s. This rapid rate of development turns out to be roughly true (e.g. Nordhaus (2001)). Working at IBM during the late 20th century, he had a front row seat to the rapid development, in both software and hardware, which has shaped the world today. My father recalls visiting IBM Havant during this period as a teen. Along with the thrill of using a computer to print his name (a novel task!), he remembers the immense scale of the computer rooms which were "probably simply doing 2 plus 2".

Later, at 16 years old, Richard Winsor began a summer internship at IBM in their cost engineering department. The department was responsible for creating 'job cards' which contained information about which parts and how many were required for each project. At this time, programming was done in 'basic', a binary language, and used paper cards to input the data where holes were punched into them to denote 1s. By feeding these cards into the computer, the total costs were calculated by the computer. This is similar to the technology developed by IBM for the NASA Apollo Space program. In this case, computer literally means machine which computes!

At the same time IBM had developed and was using PROFS (PRofessional OFfice System). This software, while extremely basic in today's terms, allowed the global company to write basic text documents and send them to its offices worldwide. This considerably sped up the transfer of information around the company and was a forerunner for modern email.

However, this type of technology was only for the few. Computers were expensive and large. The Commodore PET retailed at over £2,500 in today's money and only had 4kb memory. This meant it could store 4096 bytes which equates to 4096 characters, about a single page of text in this thesis.

It was really the 80s, when my dad started work after university, that he started to see the impact of digital technology in the ordinary office. While the office computer sat at its own station, it was now on the same floor-plate as employees, who could 'go to use the computer'. Access was increased and increasing amounts of data could be stored and accessed digitally. However, working as an accountant, Dad recalls the basic software, especially for spreadsheets. Lotus123, the office software package, had a forerunner for spreadsheet technology which could do the four basic operations using a single sheet of limited size. Anything further had to be inputted to Data-ease, a database software for adding spreadsheets.

During this decade, 'portable' computers were also introduced at the office. As part of the audit process at client offices, Dad's team would take a computer to store and process their findings. Dad remembers that the team dubbed these 'luggables' as they were large 12kg, luggage size machines and it constituted your day's exercise to 'lug' them around the city! While computers were becoming more commonplace, they were by no means convenient!

The 90s saw a rise of personal computing. One of the most significant advances was the introduction of Windows Operating System (OS) to the office. While first released in 1985, an upgrade in 1990 increased adoption of this OS. Windows represented a significant improvement to the user interface with the ability to point and click, rather than simply type commands, to switch between 'windows'. What's a window you may ask - these represented the programs and allowed more than one to be used at a time, revolutionary! The system even had basic games, like solitaire.

During this development phase, there was a constant trade-off between software and hardware. Dad recalls how frustrating and slow software could be. Turning on the computer and then going to make a cup of tea and a chat before coming back to find it still 'booting up'. As fast as hardware technology progressed - increasing computing power, decreasing sizes - software moved faster. Additionally, hardware-software compatibility was an important consideration as many pieces of software were written for specific systems so investing in expensive technology could 'lock-in' decisions for years to come.

My dad describes a turning point in technology with the introduction of Windows 95. A user-friendly interface was paired with software written specifically for the hardware - allowing far more sophisticated programs. Dad say that the inclusion of functions, such as 'sum' and 'if', to excel changed his life.

Technology was still extremely expensive and represented a significant investment by companies. Computers were subject to 'generational planning' with upper management provided the newest models and older ones cascading down the business. This contrasts with today's model where new technology tends to either directly replace the oldest and stays with that person or goes to those with the greatest computing requirement - such as the company's designers or data analysts.

This story is one man's account of how the rise of digital technology has transformed the flow of information and knowledge in the workplace. However, this is a story which many 'boomers' will recognise, from the frustratingly slow software to the mind-blowing capability of today's technology. And this progress is still going.

In this account, several core tasks of knowledge management emerge from the development of more sophisticated technology. The most prominent are knowledge storage/retrieval and knowledge sharing. To clarify, these processes are not invented by the development of new technology - human-human knowledge sharing/transfer is millennia old, with skills and information being passed along generations - rather the advances in technology aim to address barriers to effective knowledge management by introducing novel ways of performing these tasks. In exploring how technology has affected 'learning from failure', each stage from the learning cycle is considered in turn.

'Learning from failure' begins with detection and identification of failure. As noted previously, a barrier to learning at this stage is non-detection and non-identification of failure. For example, unsafe situations may not be recognised as such by personnel, or may develop away from an observable location. In these cases, sensing technology can be used to detect and identify failures. Industries have been using sensing technology for years - for example, analogue temperature sensors and alarm systems allow operators to identify when systems are operating out of range. In recent years, digital sensing technology has become more sophisticated and allows automatic identification of failure. Particularly in construction, there is a growing interest in the use of computer vision to identify real-time failures, such as real-time identification of unsafe acts - for example, Ding et al. (2018). This use of video data - which can be considered personal data - has deep ethical considerations, not least is its compliance with GDPR (General Data Protection Regulation). This is discussed further in the next sub-section.

The next step in the learning cycle - report/record - is hindered by under-reporting. People may not report/record a failure for a whole host of reasons. Accessibility to recording technology, computers or forms, is one of them. As digital capture becomes more mobile via phone apps and on-site tech, this accessibility barrier is reduced. However, introduction of digital technology could also exasperate this issue, as some people simply don't like new technology or it could make the process unnecessarily complicated. Socio-technical system optimisation theory is key here.

Storage and retrieval processes have also been totally transformed in the 21st century. Documents are stored digitally and are often now in 'cloud systems' which can be accessed anywhere in the globe. Retrieval from these systems is more efficient by using sophisticated searching algorithms. The most novel retrieval systems can search using document similarity, use document classification algorithms or by using 'question-answer' type systems, where the user can ask a question and gets the information they require - think virtual assistants.

These systems are cutting-edge and many KM strategies are still focused around investing in these store-retrieve systems. As mentioned, this 'short-circuits' the learning cycle and essentially restricts them to creating sophisticated libraries. However, organisations are no longer content for information to be simply stored for posterity; they want to be unlocking the potential business advantages such information holds.

For numerical data, systematic pipelines for data analysis exist in many business, aiming to aggregate data into meaningful dashboards and other visualisation methods which facilitate decision-making. Meanwhile, application of machine learning (ML) and artificial intelligence (AI) (next section) is growing - both automating these analysis processes and exploring advanced tasks like prediction.

However, interpreting these visualisations and identification of lessons is still a human based task. At this time, AI processes simply do not have the breadth of contextual data to project the patterns found during analysis into the real-life context.

In similar fashion, knowledge application is mainly human-centred. There are very few machines which adjust organisational procedure or make decisions without human oversight.

Knowledge transfer/sharing is also facilitated by this increase in technology. Email has transformed the flow of information around the globe. Systematic information sharing - in the form of newsletters or alerts - is often part of organisational policy, however everyday email/communication technology use is often left out of or separate to formal KM company strategies.

2.4.3 How is knowledge management different in the construction industry?

An essential text for the consideration of knowledge management in construction is *Knowledge management in construction*, Anumba, Egbu, and Carrillo (2005). Despite being over a decade old, this book contains a comprehensive guide through KM theory as it pertains to the construction industry, and provides practical insights on implementation of KM strategies in construction. By identifying key themes, such as tacit-explicit knowledge types and human-factors, the research presented explores implications of these theories in construction. Within this text, several core characteristics of the construction industry emerge as crucial in determining how knowledge management is undertaken here, as opposed to other industries. These characteristics are:

1. Project-based work
2. Workforce characteristics

Projects are unique, time-constrained and geographically distant. Ren, Deng, and Liang, 2018 found that these factors lead to a decrease in the effectiveness of knowledge management processes in construction, especially knowledge sharing and retrieval. They reported that geographical distance "may lead to gaps in languages, cultures and customs, [which] increases the communication cost and decreases the possibility of face-to-face communication". Additionally, the temporary nature of projects led to staff dispersal increasing the difficulty in knowledge management between projects. For construction, this is exasperated by the high proportion of contracted workers who move even more frequently than the project cycles.

The construction workforce is nomadic - insofar as the majority are contractors - and non-academic. This can manifest in a reluctance to use technology and learn new ICT systems at every site. While large construction programmes can enforce this as they can provide good salaries for a prolonged contract, staff could be discouraged to join smaller projects if new technology skills are required and with the current lack of skilled labour in the UK, this could lead to a lack of implementation or use of knowledge systems on project. Esmi and Ennals (2009) reported that these factors result in a reliance on tacit knowledge and human-human knowledge sharing, with limited technical support systems.

In the last decade, probably the most prominent introduction of KM strategy to construction is the implementation of 'BIM' processes.

BIM - Just another fancy store-retrieve system?

Development of BIM (Building Information Modelling) is a UK government priority to achieve "significant improvements in cost, value and carbon performance through the use of open sharable asset information" (UK BIM Alliance, 2019). BIM has been proclaimed as a step change in knowledge management for the construction industry, by providing both physical and functional information in a common data environment (CDE). From 2016, the UK government has mandated the use of Level 2 BIM (now known as PAS 1192 series) on all centrally-procured public construction contracts and the UK BIM Framework (https://ukbimframework.org/, launched in October 2019) provides detail and guidance for implementing BIM across UK construction. Additionally, BS ISO standards (BS EN ISO 19650 series) have been developed and continue to evolve as the technology develops.

The latest NBS report reflects on the changes over the last 9 years. They found, via surveys, that "In 2011, 43% of respondents had not heard of BIM. Today, awareness is almost universal, with 73% using BIM." However, the report noted that this implementation is far more predominant for large companies and projects, with a significant minority still believing that BIM is unsuitable for small construction jobs.

But, what is BIM? Building Information Modelling (BIM) refers to any system creating digital information about a physical structure - e.g. buildings or infrastructure. It is useful to use the UK BIM maturity levels to understand what BIM actually entails. While the 'Level 2' has been superseded by PAS 1192 series, it is still useful to consider these general levels. The following definitions are from Designing Buildings *BIM maturity levels* 2019:

"Level 0 – Unmanaged computer aided design (CAD) including 2D drawings, and text with paper-based or electronic exchange of information but without common standards and processes. Level 1 – Managed CAD. This may include 2D and 3D information such as visualisations. Level 1 models are not shared between project team members. Level 2 – Managed 3D environment with data attached, but created in separate discipline-based models. Level 3 – A single collaborative, online, project model with construction sequencing (4D), cost (5D) and project lifecycle information (6D)."

As mentioned, currently the UK government mandates Level 2 (PAS 1192) BIM for centrally-procured public infrastructure contracts. At minimum, this requires a 3D CAD model of the physical structure and all project data stored digitally. However, the real advances in efficiency and carbon saving are expected at BIM Level 3 - where the information is integrated and BIM can support activities such as 4D construction sequencing and engineering method selection.

However, is it just a fancy store-retrieve system? At present, this appears true for most implementations. Wang and Meng, 2019 found that knowledge capture, knowledge sharing and knowledge storage/retrieval were prominent BIM tasks. They found that these tasks are improved by BIM-supported knowledge management due to BIM's distinctive features: "object-oriented modeling, collaborative working, and digital visualization". However, when these tasks are compared to the learning cycle, analysis tasks are notably absent. It appears BIM systems are used as an advanced and more accessible[a] database, rather than as an integrated tool. NBS's report found that "BIM is still seen by many as just 3D models used

[a] On projects I have worked on, the BIM system/model is operated and accessed by a dedicated team of 'BIM Technicians/Engineers'. This did not seem to me to be more accessible than the 2D drawing database which was stored in an intranet folder system and could be searched at will. The BIM software required high specification computers and high levels of training to use - both of which actually limited the accessibility of the system. In my opinion, BIM developers need to balance technical complexity of such models with benefits. Should we be training specialists and specialist tools, or creating simpler tools? Or a tool with two modes - 'developers' and 'users'?

by designers, PMs and construction teams must get on board for BIM to be a success, it is the biggest blocker in our business." However, these observations could simply reflect the current maturity of BIM implementation, which will evolve as BIM Level 3 is developed.

Currently, analysis and knowledge from BIM seems to be gleaned by human analysis or by extracting data from BIM systems as input data into other models. These tasks are external to the BIM systems. As such, BIM is currently a sophisticated store-retrieve system. However, it still represents a huge advance in term of KM systems for construction management. BIM systems also embody an industrial priority for accessible data, presented visually, in order to increase productivity, reduce carbon emissions and decrease cost.

It would be remiss not to acknowledge the body of research for the construction industry which relates KM processes to innovation activities. For example, in just one journal (*Construction Innovation*), Walker (2016) found 203 papers published which contained the phrase "knowledge management" in a 10 year period (2005-2015). This body of research is not discussed in this literature review, returning to the assumption that 'innovation' is not 'learning from failure'.

What sources of failure knowledge exist for learning from failure on construction?

Having established that most technical KM systems in construction are developed for document storage-retrieval, this research requires identification of knowledge sources relating to failure.

The most contextualised, detailed and nuanced repository of project failure knowledge is within the people who worked on it. However complete reliance on human-human transfer can be considered ineffective, therefore how is this knowledge (and other failure data) collected in documents and construction artifacts.

Project reviews and 'lessons learnt' documents are intended to capture and transfer this knowledge and lessons between projects and teams. These reviews are infrequent and their quality relies on the manner in which they are conducted and staff support, both in terms of time and input. However, Anumba, Egbu, and Carrillo (2005) identified that these processes "are useful in consolidating the learning of the people involved in a project, but they are not very effective in transferring knowledge to non-project participants".

Construction projects also collect reports on more frequent failure documentation in the form of non-reportable incidents, safety observation/inspections and quality non-compliance reports (NCRs). This is a lower detail, high volume dataset. The variety and size of the databases are vast, and the information is often stored in a few lines of prose, an unformatted data style. These could also be used for learning. However, individually reviewing these cases is prohibitively time-consuming.

Therefore, new ways of accessing the information are required to make this suggestion viable. Use of free-text data from construction sites offers a solution to the lack of focus on small, frequent precursors - identified as a barrier to learning from failure (Stemn et al., 2018). This thesis focuses on the development of a learning process for this data, specifically addressing the lack of effective learning from incident reporting systems, by exploring methods to analyse these unstructured, text data and how to integrate insights into learning processes.

2.5 An introduction to AI (Artificial Intelligence)

Organisational learning and knowledge management processes have been transformed by modern technology. Artificial intelligence (AI) and machine learning (ML) methods have allowed automation of KM processes, facilitated advanced document retrieval and introduced advanced analysis methods. This section aims to provide readers with an appreciation for these fields and their current application in the construction industry.

To provide clarity, the taxonomy of AI is introduced before a brief exploration of AI & ML adoption in the construction industry. While key terms are introduced and explained, detailed theory of specific methods and method critique for this research are introduced in Chapter 5.

The field of Natural Language Processing (NLP) is also defined with a brief explanation of the two key methods for transforming the text into a numerical vector - an essential first step for analysing text data. The final sub-section, sub-section 2.5.4, describes the limited application of NLP in the construction industry.

The text for this section is adapted from the background sections of Baker, Hallowell, and Tixier (2020b) and Baker, Hallowell, and Tixier (2020a).

2.5.1 What are Data Science, AI and Machine Learning?

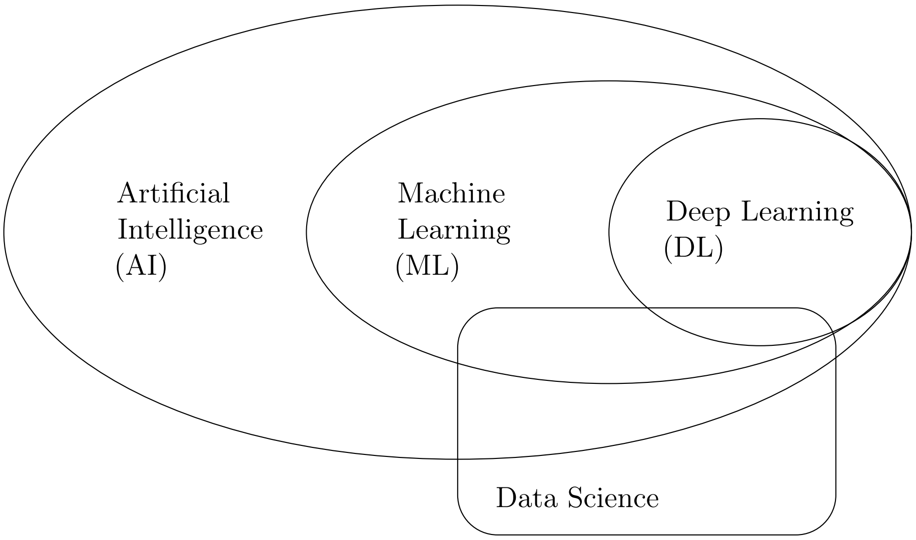


Figure 2.11: AI - Machine Learning - Data Science

Artificial intelligence (AI) describes any automated process which mimics human-like behaviour. The first computer AI systems, such as Zuse's chess program (Bauer and Wossner, 1972), rely on procedural commands, which means that they captured the intelligence of the programmer through instructions. Many AI systems still work on this basis using sophisticated versions of an "if x then y" logic. While these systems capture some human intelligence in the commands, they lack the ability to improve their task based on experience i.e. if the system carries out the task 1000 times, it is no better at performing it than the first time.

A subset of AI which originated in the 70s and 80s relies on patterns and inference from data, rather than explicit instructions, to achieve their aims and is known as machine learning (ML) (Bishop, 2006). ML describes a wide range of computer programs that implement algorithms and statistical models to carry out tasks (Hastie et al., 2005). These systems can 'learn' to carry out their task more efficiently, or to high degrees of accuracy, by using data to improve variables within the algorithms.

Recently, this has been divided again to a subset known as deep learning. Deep learning relies on 'big data' and methods like neural networks. These methods still use data to improve the task performance, however, the scale of the number of variables is magnitudes greater therefore the amount of data to 'learn' these variables is also greater.

Goodfellow, Bengio, and Courville (2016) define deep learning as machine learning methods which *"allow computers to learn complicated concepts by building them out of simpler ones"*. If represented graphically, these models have many layers, the number of which is referred to as depth. Hence if a model has multiple layers, it is referred to as *deep*. Deep learning methods typically rely on large quantities of data to train their parameters. For that reason, their increased prominence coincides with the global increase in both data availability and computational power.

Neural networks are the most common collection of deep learning architectures and often the terms are used interchangeably. While they were initially developed as early as 1940s, these simplistic early networks have undergone radical developments and increases in levels of sophistication, achieving record pattern recognition levels since the 1990s (Schmidhuber, 2015; LeCun et al., 1998).

Data science is a field of work which uses these methods, but also relies on other knowledge such as interpretation, data bias etc. This research, concerned with the selection and application of AI methods to solve a defined problem, lies within the realms of data science. These relationships are illustrated in Figure 2.11.

2.5.2 How is machine learning currently used in the construction industry?

Machine learning (ML) in construction has been developed significantly since 1991 when Moselhi, Hegazy, and Fazio (1991) first discussed the potential of neural networks in construction engineering and management. Early examples of ML in construction include applications such as that of Skibniewski, Arciszewski, and Lueprasert (1997), where the AQ15 algorithm was applied to automatically learn the mapping between constructability (poor, good, excellent) and 7 predictors from a collection of 31 training examples; and that of Soibelman and Kim (2002) who applied decision trees and neural networks to a construction management database to identify the causes of delays.

Many subsequent prediction applications applied support vector machines (SVMs), owing to their consistently high accuracy. These applications include Lam, Palaneeswaran, and Yu (2009), who accurately forecasted contractor prequalification using input variables such as financial strength and current workload; Cheng et al. (2010), who estimated building cost and loss risk from ten input variables; and Son, Kim, and Kim (2011), who detected concrete structural components in color images from actual construction sites.

In the last 5 years, use of ML in construction has become far more widespread and the methods and applications used are far more diverse. In addition to classic prediction tasks, more nuanced applications have emerged. Some interesting examples include construction equipment activity recognition (Akhavian and Behzadan, 2015), and productivity and ergonomic assessment (Nath and Behzadan, 2017).

Machine learning for construction safety

In terms of using ML to learn or gain insights from failure events, the most developed field of construction research is ML application to construction safety.

Before 1995, research was heavily invested in the analysis of lagging statistics Zhou, Goh, and Li (2015). The aim of such studies, e.g., Hubbard and Neil (1985) and Salminen (1995), was to observe trends in accident numbers and postulate correlations with a limited number of circumstantial factors to suggest future safety measures or research avenues. At the same time, statistics concerning safety incidents and their associated cost were used to create financial motivation for safety research, e.g. Koehn and Musser (1983). Neither of these applications attempted to empirically forecast future trends or safety events, but rather examined the current state and postulated positive actions towards reducing incident rates.

Recently, research regarding pure prediction of construction safety outcomes from descriptors of the work and the work environment has emerged. A survey by Hallowell, Bhandari, and Alruqi (2019) recognised two studies: Tixier et al. (2016a) and Esmaeili, Hallowell, and Rajagopalan (2015). Further publications identified in this domain are Kang and Ryu (2019), Sarkar et al. (2019b), Sarkar et al. (2019a), and Baker, Hallowell, and Tixier (2020b). However, in all these pieces of work except Tixier et al. (2016a) and Baker, Hallowell, and Tixier (2020b), some of the input variables are outcomes. Such variables cannot be considered valid predictors as they are not observable before accident occurrence. E.g., in Esmaeili, Hallowell, and Rajagopalan (2015), *structure collapse* and *falling from roof*, two outcomes, are used as attributes. The attributes of Sarkar et al. (2019b) and Sarkar et al. (2019a) also include two outcomes: incident type and injury type. Finally, Kang and Ryu (2019) rely on accident type and injury type too, but also on body part injured and accident location. All of these variables, again, are outcomes, not predictors. These papers, therefore, have applied ML methods, but failed to correctly consider the context and implications of the method to the construction context - professionals require methods of prediction which do not rely on the outcome as an input.

Another relevant study is Poh, Ubeynarayana, and Goh (2018). In this study, the authors rely on 13 project management and safety-related leading indicators from monthly inspection data (before incident occurrence) to make severity forecasts.

2.5.3 What is Natural Language Processing (NLP)?

Natural Language Processing (NLP), also known as computational linguistics, is a rapidly developing field dealing with the computer analysis of both written and spoken human language. It is acknowledged to be an interdisciplinary field, using concepts from linguistics as well as computer science, statistics, and machine learning in general. As well as applications in speech recognition and machine translation, NLP has gained interest in text retrieval and automated content analysis - both of which can be described as knowledge management tasks.

Transforming unstructured free-text data into a structured representation is a key preliminary task in many NLP applications. NLP also has many other areas of research and theory, relevant for different applications, which will not be detailed in this research.

Once a structured representation of the text has been achieved, further analysis and machine learning tasks can be performed, such as text retrieval or classification. Classification tasks using machine learning classifiers can be performed for both binary and multi-class classification. Meanwhile, vector similarity can be mathematically evaluated to rate the similarity of a vector against another in the corpora allowing similar documents to be retrieved.

While early researchers focused on writing lexical rules which computers could follow, this was found in most reports to be unwieldy due to word ambiguity and grammatical complexity, giving rise to the popularity of empirical language models in the late 1980s (Hirschberg and Manning, 2015; Katz, 1987).

It is not until recently that, following the advent of distributed word representations, e.g. Bengio et al. (2003), Mikolov et al. (2013a), and Mikolov et al. (2013b), deep learning architectures have been developed for NLP tasks such as natural language understanding and machine translation (with great success) (Kim, 2014; Luong, Pham, and Manning, 2015).

This subsection briefly outlines these two models before current use in construction is explored.

Vector space representation AKA ‘Bag-of-Words’ (BoW)

Empirical representation, based on the *Bag-of-Words* (BoW) representation (also known as the *vector space* representation), have dominated the research space since 1980s due to their notable results when trained on large datasets (Hirschberg and Manning, 2015). These representations are based on the numerical frequency of unique ‘tokens’ contained within the training vocabulary. ‘Tokens’ generally include words but also may also include punctuation or numbers. The resultant representation is a very long, sparse vector.

With BoW, a given document is represented as a vocabulary-size vector that has zeroes everywhere except for the dimensions corresponding to the tokens in the document. The vocabulary is made of all the unique tokens in the preprocessed training set. Depending on preprocessing, tokens may include words, phrases, punctuation marks, numbers, codes, etc.

BoW ignores word similarity and word order. For example, “hammer fell on worker” and “worker fell on hammer” have the same representation, and “hammer” and “tool” are not considered more similar than “hammer” and “worker”, as all dimensions of the vector space are orthogonal. This restricts the semantic meaning which can be gained from such representations. To capture word order locally, combinations of tokens (i.e., phrases), formally known as n-grams, may be used instead of single tokens. But doing so makes the vector space become so large and sparse that it makes it hard to fit any model, a problem colloquially known as the *curse of dimensionality.* In practice, it is rarely possible to use n-grams of order greater than 4 or 5.

Finally, some syntactic information may be captured by creating different dimensions for the different *part-of-speech* tags of a given unigram (noun, proper noun, adjective, verb, etc.), but this has the same adverse effects on the dimensionality of the space as that previously mentioned.

Word embedded representation AKA 'word embeddings'

New ways of representing textual data are based on *embeddings*, also known as *word vectors* or *distributed word representations*. With word embeddings, each word in the vocabulary is represented as a small, dense vector, in a space of shared concepts. To derive a representation for a document, the vectors of its words are combined, either simply through averaging or concatenation, or through more sophisticated operations (neural networks). One should note that character or subword embeddings are sometimes used, e.g. Zhang, Zhao, and LeCun (2015) and Bojanowski et al. (2017), with the main benefit of providing robustness to out-of-vocabulary words and typographical errors. However, the word is the most common granularity level.

Unlike the long and sparse BoW vectors, word vectors are short (typically 100-500 entries), dense, and real-valued. The dimensions of the embedding space are shared latent features, so that after training, meaningful semantic and syntactic similarities, and other linguistic regularities, are captured. For instance, Tixier, Vazirgiannis, and Hallowell (2016) applied the unsupervised `word2vec` Mikolov et al. (2013a) model to a large corpus of construction-related text. In the final embedding space, a constant linear translation was found to link body parts (tendon, brain) to sustained injuries (tendonitis, concussion), and another one to link tools and equipment (grinder, chisel) to the corresponding material (metal, wood).

Deep learning architectures are fed the sequence of word vectors of the input document and pass them through their layers. Each layer computes a higher-level, more abstract representation of the input text by performing operations (e.g. convolutional, recurrent) on top of the output of the previous layer, until a single vector representing the entire input document is obtained.

Then, depending on the task, one may add a few specific final layers (e.g. dense, sigmoid, softmax for regression or classification), a decoder (sequence-to-sequence setting for translation or summarization), or combine two encoders via a meta-architecture (e.g. siamese or triplet configuration for textual similarity Shang et al., 2019).

The word vectors are not necessarily initialized at random, like the other parameters of the network. It is actually advantageous to pre-train them in an unsupervised way. Then, word vectors can either be fine-tuned or kept frozen during training. When pre-training is conducted on an external, typically large corpus of unannotated raw text, the approach is known as *transfer learning*. To this purpose, unsupervised, shallow models such as `word2vec` (Mikolov et al., 2013a) or `GloVe` (Pennington, Socher, and Manning, 2014) can be applied to big corpora like entire Wikipedia dumps or parts of the Internet[3]. `ELMo` (Peters et al., 2018) and `BERT` (Devlin et al., 2018) have also made great strides recently, by showing that it was possible to transfer not only the word vectors but the entire model. After pre-training, the model (or simply its internal representations in the case of `ELMo`) are used in a supervised way to solve some downstream task, for example, sentiment analysis or named entity recognition. `ELMo` and `BERT` have brought great improvement to many natural language understanding tasks.

However, to date, no significant advantage to accuracy has been achieved using deep learning methods for classification of safety event descriptions, despite the increase in complexity.

Feature representation

Another way of representing text is through features learnt via other methods. Each dimension in the vector represents some meaningful feature. For example, a word can be complemented by its stem, PoS, word shape etc. This is akin to feature engineering for ML.

These features could also relate to the text as a whole, rather than the individual word. In this way, a list of key words could be considered a feature representation for the piece of text -

[3] e.g., `https://code.google.com/archive/p/word2vec/` (under section "pre-trained word and phrase vectors"), `https://nlp.stanford.edu/projects/glove/`

although to convert this list into a form useable for ML, conversion to a numerical vector probably via BoW would be required.

2.5.4 How is NLP currently used in the construction industry?

Global interest has grown in applying NLP for comprehension and analysis of construction documents. However, nearly all examples found use the BoW representation, losing the semantic relationships between words and ignoring word order.

Existing literature reveals a number of different natural language tasks performed in the construction sector, including classification and retrieval of documents. Caldas and Soibelman (2003), Goh and Ubeynarayana (2017), and Zhang et al. (2019) compare machine learning classifiers using BoW inputs and all find that Support Vector Machines (SVM) [4] result in the highest accuracy. In other classification tasks, Chokor et al. (2016) and Marzouk and Enaba (2019) elected to cluster the BoW vectors. For two document retrieval tasks, the researchers used vector similarity to identify the most relevant reports (Yu and Hsu, 2013; Zou, Kiviniemi, and Jones, 2017).

Some studies attempted to adjust for the shortcomings of the BoW representation. Zou, Kiviniemi, and Jones (2017) and Kim and Chi (2019) attempted to recapture some semantic relations by implementing thesaurus relations into their BoW vectors; however, this required the use of construction specific dictionaries to supplement thesaurus definitions from general lexicons due to the specificity of construction language. Williams and Gong (2014) incorporated bigrams into their text representation in order to capture some of the local word order; however, they found that higher level word groupings were unable to significantly increase the accuracy of the predictions. Finally, Tixier, Vazirgiannis, and Hallowell (2016) used a Wasserstein distance in the word embedding space for injury report retrieval and classification (with the k-nearest neighbors algorithm, for classification).

Meanwhile, Tixier et al. (2016a) and Tixier, Hallowell, and Rajagopalan (2017) extracted 81 fundamental attributes (or precursors) from injury reports using a tool based on an entirely hand-written lexicon and set of rules, reported in Tixier et al. (2016b). This allowed them respectively to predict safety outcomes with good accuracy, and to identify interesting combinations of attributes, coined as "safety clashes". However, the development of the tool was resource intensive, both in terms of time and human-input requirement.

Only two papers identified, both published since the start of this research, experiment with embedded word representation for classification tasks. These are Baker, Hallowell, and Tixier (2020a) and Zhong et al. (2020). They report a slight improvement in classification results over using BoW representations, but nothing significant. In a different task, Sun et al. (2020) used deep learning representation methods to help visualise key word connections extracted from quality records written in Chinese.

There has been limited experimentation with deep learning methods for text data in related subject areas, such as Chung (2018) who applies recurrent neural networks with LSTM units to perform named entity recognition on bridge inspection reports. This is not construction text; however, future research in these related civil engineering fields could yield insights relevant to the construction industry.

These works demonstrate the potential of NLP in the construction domain. During the course of this research, there has been a surge of interest in application of NLP, especially in the classification of safety incident descriptions. These are explored in more detail in Chapter 5 where critical analysis of this literature informs the AI method selection, rather than the problem domain.

[4] Detailed explanation of this method is included in Chapter 5

2.6 Summary

To weave together the context and literature explored here, recall Figure 2.1. The first step aimed to explore the concepts of 'failure' and 'organisational learning' in construction as well as detailing the current state of 'learning from failure'.

It was found that there is a lack of agreement and foundation to the definition of failure in construction, especially when limiting the literature to a UK context. In particular, failure seemed to be often defined as the absence of success. Success literature, including topics such as identification of success criteria and factors and assessment methods, appear to be directly applied to failure. However, there was no evidence found to support the belief that the criteria and factors to achieve success are equally important to avoid failure. In fact, the psychological, cultural and human factors literature suggest that this assumption could be incorrect or at least require investigation. As this understanding is essential for this research, it must be investigated further before research into suitable 'learning from failure' systems can begin. This literature gap led to the qualitative investigation presented in Chapter 4.

In considering organisational learning, it was identified early into this research that, in order to properly inform the investigation, more information would be required than could be gleaned from the existing studies. Therefore, the literature presented is supplemented with findings from the preliminary investigation, presented in Chapter 4. However, from the literature which does exist, it appears that the construction industry emphasises building personnel competence, with large organisational learning mainly occurring through adjustment of standards following large, infrequent reviews.

In relation to learning from failures, this means that the construction industry learns from catastrophic events or case studies, and relies on individuals' interventions to attempt to incorporate and communicate learning from smaller events - either through alerts or stories. There appears to be a lack of formal processes to analyse failure data beyond lagging statistics and also to connect the communication of failures to implementation of change.

Exploration of these concepts led to consideration of knowledge management - "the conscious organizational and managerial practices intended to achieve organizational goals through efficient and effective management of the firm's knowledge resources" (Inkinen, 2016). This found that the rise in digital technology has drastically changed the processes of knowledge management. In particular, data analytics has allowed advanced analysis and visualisation in many domains.

Data science methods, in particular NLP (natural language processing) were shown to hold the potential to unlock the knowledge trapped in failure data sets. However, the construction industry is historically cautious with adopting new technology, therefore any new methods must be contextually suitable and take this into account.

In conclusion, the existing literature reveals the potential of unstructured failure data to delivering insights for the construction industry. However, a lack of underlying understanding about the concept of failure and learning from failure in construction currently hinders application and further exploration. This research first addresses this literature gap through a qualitative analysis in Section 2.3.2 then explores how novel data science methods can facilitate and be implemented for learning from failure in the construction industry.

Chapter 3

Methodology

"If there's no such thing as objectivity then there's no such thing as measurement which means that empiricism is meaningless"

Bones Episode 9 Season 6

During this research process, I particularly enjoyed engaging with the philosophical discussion surrounding research methodology. This quote from the TV show 'Bones' is particularly apt to this discussion. Dr Temperance Brennan, the US crime drama's protagonist, is a forensic anthropologist whose work involves collecting evidence from skeletal remains. This laboratory-based work is based firmly in positivism, where the truth can be measured and would remain the same for any researcher. As a result, she has an extremely sceptical view on subjective measurement - such as required by any non-lab-based research. In this Chapter, I refute her view concerning the meaningless nature of subjective measurement and establish the philosophical basis of this research.

3.1 Methodology

In this work, methodology is defined as the philosophical foundation upon which the researcher undertakes research. This includes their philosophical assumptions about what constitutes as reality or '*the truth*' as well as their approach to research design and method selection. In any type of research, it is important to have an appreciation of underlying methodology in order to inform decisions, such as research method selection, and to be aware of potential research biases. This subsection first introduces some key methodological concepts, then applies these to inform the research design.

To develop a philosophical view of one's research, Easterby-Smith, Thorpe, and Jackson (2012) suggest the analogy of a tree, working outwards to construct a solid 'trunk' on which to base one's research. Meanwhile, Saunders, Lewis, and Thornhill (2009) use the analogy of an onion, working inwards to peal back the layers and focus one's research. However, they both agree that the key path to consider when setting the foundation of any research problem is: ontology, epistemology, methodology (outlining research design decisions which differ slightly in content for each analogy) and methods/techniques.

The *ontology*, describing the philosophical nature of reality, and *epistemology*, describing assumptions on the ability to discover or inquire about this reality, are closely coupled and often conflated. In considering these philosophical premises, Easterby-Smith, Thorpe, and Jackson (2012) present a separate scale for each, containing key philosophical viewpoints and their definitions, before stacking these scales to show their relationship. Further philosophies can then be placed along these scales. As presented in Figure 3.1, they first present four different ontologies which range from *realism* to *nominalism.* This is followed by a binary epistemological comparison of *positivism*, where researchers are independent from the process and hypothesis they are testing, vs *constructivism* (aka social constructivism), where actors (including researchers, research participants and society at large) interact with the phenomena being investigated. They then note that these two scales, for ontology and epistemology respectively, are closely related as the definition of *'the truth'* leads strongly to the nature of investigation available. This also has implications for the research design (which they call methodology), as seen in Figure 3.2.

A key consideration in forming a methodological view is the approach the research will take: inductive or deductive (Saunders, Lewis, and Thornhill, 2009). These are opposite sides of the same process. Deductive research works from a theory to form a hypothesis and then aims to prove (or disprove) this via application of research methods to gather data. This type of research is generally described by principles on the left hand side of Table 3.2, where data is quantitative and investigations can be described as theory-driven. On the other hand, inductive research works from observations, or data, which are collected by the researcher who then forms a hypothesis from this which can be further developed into a theory. This generally describes research on the

Table 3.1: Illustration of Ontology Scale (Easterby-Smith, Thorpe, and Jackson, 2012)

Ontology	Realism	Internal Realism	Relativism	Nominalism
Truth	Single truth.	Truth exists, but it is obscure.	There are many 'truths'.	There is no truth.
Facts	Facts exist and can be revealed.	Facts are concrete, but cannot be accessed directly.	Facts depend on viewpoint of the observer.	Facts are all human creations.

Table 3.2: Combined Illustration of Ontology, Epistemology and Research Design (Easterby-Smith, Thorpe, and Jackson, 2012)

Ontology	Realism	Internal Realism	Relativism	Nominalism
Epistemology Methodology	Strong Positivism	Positivism	Constructionism	Strong Constructionism
Aims	Discovery	Exposure	Convergence	Invention
Starting Points	Hypothesis	Propositions	Questions	Critique
Designs	Experiment	Large surveys, multi-cases	Cases and surveys	Engagement and reflexivity
Data types	Numbers and facts	Numbers and words	Words and numbers	Discourse and experiences
Analysis/ interpretation	Verification/ falsification	Correlation and regression	Triangulation and comparison	Sense-making and understanding
Outcomes	Confirmation of theories	Theory testing and generation	Theory generation	New Insights and action

right hand side of Table 3.2. In recent history, this is dominated by qualitative research and can be described as data-driven. It is interesting to note that the rise in 'big data' has the potential to change this. Hey, Tansley, and Tolle (2009) describe in their book the "Fourth Paradigm of Science" which they propose is the fourth generation of quantitative research where discovery in large datasets drives theory generation and provides insights. I elaborate on this line of thought in discussion of the method selection in Chapter 5.

For physical sciences, consideration of methodology is often trivial, or implicit, as '*the truth*' is generally accepted as a phenomenon which can be **proven** (or verified) by **repeatable** experiments or investigations where the role of the researcher is minimal, i.e. it would not matter who was performing the experiment or research, the results would be the same. This places most investigations firmly towards the left-hand side of Easterby-Smith, Thorpe, and Jackson (2012)'s illustration, with strong positivist assumptions. However, social sciences must more closely examine the formation of their methodology and philosophies as the involvement of the researcher in their own research, the subjects of the research and their interactions exert pressures upon the investigations and therefore the outcomes they produce. Saunders, Lewis, and Thornhill (2009) consider the role of the researcher's own values and opinions to be a distinct component for consideration, which is referred to as the *axiology*. For Easterby-Smith, Thorpe, and Jackson (2012), this is implicit in the choice of *ontology* and *epistemology* and not considered separately. It should be noted that *axiology* was recently adopted to cover matters concerning the philosophy of values which, according to Hiles (2008), "covers a wide area of critical analysis and debate that includes truth, utility, goodness, beauty, right conduct, and obligation."

There exist many other philosophical views, with differing levels of nuance, upon which researchers base their assumptions and research design decisions. To list or describe them all would not add value here; however, a few epistemological theories relevant for consideration in this research are included below:

- Positivism. As previously described, this epistemology stems from the acceptance that there exists a single, objective reality which can be captured and revealed (the realism ontology). According to Paley (2008), the iconic depiction of this philosophy stems from two key fundamental commitments: "to empiricism (i.e. there is knowledge only from experience) and to logical analysis, by means of which philosophical problems and paradoxes would be resolved and the structure of scientific theory made clear." He also acknowledges that rarely, if ever, does any research accept this package of beliefs in its entirety, rather leaning towards modified versions.
- Post-positivism. This philosophy is considered a critique of positivism as researchers attempt to capture the impact their own values and background has on the research. While they generally maintain that there exists an objective *'truth'* (realism ontology), there is also acceptance that the researcher can affect their data and analysis, and that these effects should be captured and taken into account. This philosophy therefore still falls under '*internal realism*' on Easterby-Smith, Thorpe, and Jackson (2012)'s scale (see Table 3.1), where an objective truth exists but this truth is obscured as research methods are flawed. However, the philosophy borrows understanding from theories which adopt a relativism ontology, such as constructivism (Fox, 2008).
- Interpretivism, also known as anti-positivism. Stemming from post-positivism, this theory's fundamental distinction is the belief that research on human beings by human beings cannot yield objective results, and therefore they look for meaning in the subjective results. This school of thought generally falls under a more relativism ontology and steps away from theory testing, aiming instead to generate new theory.
- Constructivism. This epistemology rejects the existence of an objective reality and believes that individuals construct knowledge (or *facts*) through social experience (Costantino, 2008). This is sometimes used interchangeably with social constructionism; Gergen and Gergen (2008) explain, however, that social constructionism considers knowledge generation as a product of human relationships and interaction, while constructivism considers knowledge generation via sense-making in the individual's mind.

Another important concept for this thesis is the philosophy of pragmatism. Pragmatism is not strictly an epistemological theory. McCaslin (2008) explains that "*pragmatism holds that truth is found in "what works," and that truth is relative to the current situation*". This is step removed from *ontology* and *epistemology* and acts to weaken the links previously set, such as the overlaying of the scales by Easterby-Smith, Thorpe, and Jackson (2012), as research undertaken employing a pragmatism philosophy will pick the most *useful* definitions. Since its introduction, pragmatism has been heavily critiqued as being too undefined and that, in attempting to harmonise subjective and objective views, it degrades both. However, it is included due to its implication for methodology. When applying pragmatism, the research begins with the formation of the statement of the problem - no research design decisions are undertaken before this point. This is extremely relevant for many engineering research projects which aim to have direct practical impact (note *practical* and *pragmatic* have the same Greek root, *pragma*, meaning action). In fact, in the foreword to a report documenting a two-part seminar exploring the philosophy of engineering, Dr Keith Guy (the Chair of the Royal Academy of Engineering's philosophy of engineering steering group at the time) remarked that "no engineer embarks on a project unless there is an end purpose for what they are working on" (RAEng, 2010).

Therefore, in designing this research, I adopted a pragmatic approach. This is in-line with the pre-generation of the problem statement(s) and I treat this philosophy as a lens which helped to inform decisions going forward. It was therefore appropriate for each part of the research to consider each of the research questions to help inform and narrow the methodological decisions. This process is in the next section (Section 3.2) for the overall RQs and in Section 4.2 for the sub-questions for the qualitative exploration of learning from failure in the construction industry.

3.2 Revisiting the research questions

It is sensible, in light of the literature, to revisit the original research questions and objectives to assess whether these need revision. This sub-section considers each of the four research questions originally proposed, and their suggested objectives, taking into account both the literature available and the methodological lens of pragmatism i.e. 'what is the most *useful* way of phrasing and examining this aspect?'

The initial framing (from the beginning of this research journey) of these research questions is captured here alongside the initial proposed objectives to answer these questions, before they are re-examined. Research Question 3 (in **bold**) embodies the majority of the research presented in this thesis.

1. How does the construction industry currently learn from failure?
   - Literature search exploring learning from failure in the construction industry, comparing to norms and best practice in other industries
   - Undertake a qualitative investigation (This objective was added after having not found cohesive picture in the literature)
2. What recent AI and data science methods have been used in the construction industry, and what other methods exist?
   - Literature search
3. **How can novel data science methods facilitate knowledge discovery from failure data?**
   - Collect text-based failure data
   - Investigate methods to convert unstructured failure data into structured forms
   - Compare the usefulness of different knowledge discovery and Natural Language Processing (NLP) models
4. How is best to implement this type of learning into systematic processes for the construction industry?
   - Consolidate and discuss these findings in relation to application for the construction industry

Research Question 1: How does the construction industry currently learn from failure?

This question aimed to uncover the current processes and human factor aspects of learning from failure in the construction industry. This is important to the research project due to its implication for data bias and appropriate analysis design. It was initially anticipated that this information could be found in existing literature. While some literature was found to contain nuggets of information relating to learning from failure processes in the industry, the majority revolves around success, with failure implied as the opposite, and there was a distinct lack of research concerning learning from failure itself. This literature gap was also identified by Lundberg, Lidelöw, and Engström (2017), a year after this research commenced. It therefore emerged early into this research that a more thorough preliminary investigation to explore this question was required. In consideration of the question nature, it was clear that an inductive research methodology would be required and therefore a social science paradigm would be appropriate. The methodology behind this investigation is discussed further in Section 4.2.

Research Question 2: What recent AI and data science methods have been used in the construction industry, and what other methods exist?

During the preliminary review of knowledge management literature, it was apparent that application of new technology, especially data science methods, was key in developing and delivering state-of-the-art systems. This research question emerged to explore the up-take of AI and data science in the construction industry and compare this to 'exemplar' knowledge systems in other industries. It was appropriate to explore this question using existing literature, therefore, this question was covered in the literature review in the previous chapter. In terms of methodological standpoint, the literature review was conducted with the same pragmatic lens as is used in the rest of the research. That is to critique the existing literature on 'what is useful'.

Research Question 3: How can novel data science methods facilitate knowledge discovery from text-based failure data?

This question aimed to investigate the application of existing data science methods into the context of learning from failure in construction. Upon reflection, I wished to assess the usefulness of different Natural Language Processing + ML models to discover patterns and trends in the failure data. This knowledge can then be implemented into organisational learning within the construction industry. In assessing 'what is useful', model accuracy and output structure will be assessed. This research is deductive - working from an initial proposition to test it - which indicates that a more positivistic epistemology would be most useful for this investigation. Here, the desired outcome of this investigation is theory testing, not theory confirmation. Adopting an internal realism ontology, as seen in 3.2, is therefore the most suitable standpoint for this investigation, again suggesting positivism as a suitable epistemology.

In considering the nuances of positivism, strong positivism was rejected as it is most associated with realism - the adoption of a single, identifiable truth - which is unsuitable for this investigation as there is no anticipated 'single truth'. In addition, this research accepts that my own values as a researcher will affect the research - an example of this will be seen in selection of the method(s) to investigate. This foundation leads to select post-positivism as the most representative epistemology.

The current question is not suitable as 'how' is much more suited to inductive research. A more suitable phrasing of this question is: 'Which NLP + ML model best facilitates knowledge discovery from text-based failure data?'

This question also better allows logical identification of the objectives. The first objective is unchanged - data collection. Then, 'Investigate methods to convert unstructured failure data into structured forms' is revised to 'Identify methods'. Following this, 'Compare the usefulness of different knowledge discovery and Natural Language Processing (NLP) models' is rephrased as 'Test the accuracy of the selected methods'. Finally, a new objective is added 'identify/develop suitable knowledge discovery models'.

Research Question 4: How is best to implement this type of learning into systematic processes for the construction industry?

Upon reflection, this question aims to form recommendations to industry by considering application of the data analysis results (RQ3) into the context established in the inductive investigation of RQ1. This aims to generate a theory or framework of theory. The most useful theories to consider are therefore interpretivism or constructivism. In this case, constructivism is most appropriate.

- Consolidate and discuss these findings in relation to application for the construction industry

In summary, by applying a pragmatic lens to this research, different methodological stances have been identified for different sections of this research. This is known as mixed or multi-method research, as defined further in Section 3.3. To restate, RQ3 (in **bold**) embodies the focus of the research presented in this thesis. The final research questions (RQs), with associated objectives, are:

1. How does the construction industry currently learn from failure?
    - Literature search exploring learning from failure in the construction industry, comparing to norms and best practice in other industries
    - Undertake a qualitative investigation
2. What recent AI and data science methods have been used in the construction industry, and what other methods exist?
    - Literature search
3. **Which Natural Language Processing (NLP) + Machine Learning (ML) model best facilitates knowledge discovery from text-based failure data?**
    - Collect text-based failure data
    - Identify methods to convert unstructured failure data into structured forms
    - Test the accuracy of different NLP + ML models for text-based failure data
    - Identify/develop suitable knowledge discovery models
4. How is best to implement this type of learning into systematic processes for the construction industry?
    - Consolidate and discuss these findings in relation to application for the construction industry

3.3 Multi-method approach

The previous section identified several different methods, relying on different methodological philosophies, to achieve the aim of this research. This is not unusual for pragmatic research. In fact, Johnston (2012) state that "*the primary philosophy of mixed methods is pragmatism.*" While they refer to mixed methods as a general term to encompass any investigation which employs more than one research method, this research distinguishes between the mixed and multi method in the same manner as Pat Bazeley's response in Johnston (2012)'s paper:

> Multimethod research is when different approaches or methods are used in parallel or sequence but are not integrated until inferences are being made. Mixed methods research involves the use of more than one approach to or method of design, data collection or data analysis within a single program of study, with integration of the different approaches or methods occurring during the program of study, and not just at its concluding point.

This research uses different methods in sequence, however, does not integrate these methods instead using the results of the first to inform decisions for the second. There is no on-going combination of data collection or analysis. An example of this could be when a researcher investigates a topic both quantitatively and qualitatively using two different survey question types and analyses them in parallel.

This distinction is helpful in understanding the logic of the remainder of the thesis. Chapter 4 is essentially a self-contained piece of research exploring learning from failure in the construction industry. This investigation compensates for the lack/weakness of previous literature on this topic, essential for forming the assumptions and informing the interpretation of the rest of the thesis.

Chapter 4

Understanding learning from failure in the construction industry: A qualitative thematic analysis

"*Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?*"

T.S.Elliot's 'Choruses for "The Rock"

This chapter outlines an initial qualitative investigation to explore the concept of 'failure' in the context of the construction industry. Having spent the previous seven years of my life based very firmly in the positivism camp of physical science, the investigations and research contained within this chapter were a huge step outside of my normal zone. Very soon into my initial literature search, it became apparent that the background knowledge and understanding upon which to base assumptions and decisions to answer my research question did not exist in a cohesive form. Therefore, this chapter is a detailed account of my journey into the social science research methodologies and methods that were necessary to form the fundamental understanding upon which I could build my research.

4.1 Why conduct this initial investigation?

Any research must be based upon a firm foundation of context and background understanding. As outlined in Chapter 2, descriptions and definitions of the main concepts, along with their academic investigation, can be found in previous literature. However, there did not exist a cohesive study or collection of studies which created a holistic consideration of failure and learning processes in a construction industry context. A majority of the literature found viewed these topics from the perspective of success and simply implies that failure is the opposite.

In order to ascertain a solid contextualisation and determine potential assumptions and limitations of this research, it was necessary to undertake an investigation to determine the current 'learning from failure' situation in the construction industry. The questions posed were formulated from gaps or refinement required in the existing literature.

This investigation responds to the first research question: "How does the construction industry currently learn from failure?" This is broken down further into sub-questions:

1. What is defined as 'failure' in a modern construction project?
2. How is failure information captured and what are the barriers?
3. What happens to these data?
4. What learning from failure processes currently exist?

These questions are revisited and revised considering the adopted methodology. A suitable method is then outlined. This section then provides details on the development of the data collection and analysis. The thematic analysis of collected interview data leads to a narrative which addresses the sub-questions posed here. From this, insights for learning from failure are generated, both for this further research and for immediate impact in the construction industry.

4.2 Methodology and Method

4.2.1 Methodology

As mentioned in Chapter 3, this research followed a pragmatic methodology. For each part of the investigation, examination of the sub-research questions through this pragmatic lens informed the philosophical definitions and narrowed the methodological decisions. This ensured the correct methods were adopted. This process was repeated for the main body of the research.

Sub-question (1): what is defined as 'failure' in a modern construction project?

When considering this question, I reflected that what I really want to achieve is understanding about what construction employees (the social actors) consider to be a failure (the social phenomenon) and how their understanding of failure interacts with the activities performed as part of their jobs on the construction site. 'Failure' is therefore a social phenomenon which the construction professionals themselves define and then interact with. This falls firmly within the relativism ontological view with a strong suggestion for use of constructionism principles. There is no suggestion that nominalism would be a useful viewpoint as rejection of any objective truth would undermine the usefulness of the insights. While it could also be argued that the definition of 'failure' is an independent social phenomenon dictated by the company, and therefore objectivism may be applicable, my interest lay with the definition of failures which the individuals held as I felt this would yield the most useful outcomes. This question was therefore refined to (1): what does 'failure' mean to different members of the construction industry?

Sub-question (2) - (4): How is information about failures captured and what are the barriers? What happens to these data? What learning from failure processes currently exist?

These three questions are grouped for this discussion as they could all almost be considered positivist type questions; an investigation of whether something, or some type, of process physically exists or not (i.e. 'Does process X exist?'). However, as explained in Section 2.4, information capture processes are tightly coupled with human behaviour so not only are these processes a human construct but also the process itself affects the behaviour of those using it. It is tempting to say that these questions lean towards an internal realism philosophy, as defined in Table 3.1 where a single truth (in this case, the existence of a process) can be discovered; however, does a process exist when it has been set up or does it exist when people use it? To a company, the 'truth' of the existence of a process may lie in its formal inception, while an employee or manager may only consider a process to exist if it is used, and to only exist in the manner in which it is used regardless of its intended purpose. I considered these nuances extremely important during this investigation. Therefore, the most useful ontological definition is relativism. Additionally, the most useful and interesting insights lie in the meaning the construction employees (the social actors) place upon these processes and, therefore, constructionism principles are the most appropriate for this investigation. In order to better access this information the final question was rephrased as :'(4) How is learning from these mistakes/failures currently implemented? Do these methods work?'

Equally, it is important to note that none of these questions could be formed into a theory or hypothesis to be tested. They direct the researcher (me) towards inductive approaches, more suited to qualitative methods. In qualitative research, the role of the researcher is often discussed in relation to his or her impact on the research being carried out (Silverman, 2013). Preconceived notions about the possible, or probable, answers to these questions could affect the research and bias the results. I had the unusual position as a researcher where I was coming into an area I had very little academic background knowledge about. As such, I seized the opportunity to design the research prior to performing the in-depth literature review; I only reviewed enough to appreciate the knowledge gap in the construction industry. This removed some unintentional bias which I may have induced as a researcher. Previous knowledge of the theories may have encouraged confirmation bias. I should be aware that my previous knowledge of the construction industry may have caused me to look for the 'known' (to me) answers rather than seeking the entire picture. This is considered further during method selection.

To conclude, in discussing the appropriate methodology to approach these sub-questions, I applied a pragmatic lens where I decided that the most useful philosophical approach would be to adopt a relativism ontology and use constructionism principles to inform the methodology and research method decisions. I also acknowledged the role I, as the researcher, may have of the research which follows and take this into consideration throughout the process.

4.2.2 Qualitative method

Having formed the sub-questions and using the selected methodology, a suitable research method was required. To provide the most useful and insightful outcome, the method selected should allow an in-depth examination of the features behind learning processes from failure and associated attitudes. This indicated an inductive method which allowed a level of discourse. Some common data collection methods considered were:

- Ethnography.

  Ethnography is a longitudinal method with a strong constructionism philosophy where the researcher embeds themselves into the situation or organisation which they wish to investigate and experience first-hand the processes, attitudes and culture. It is time consuming but can create an extremely detailed account, enriched with personal reflections and encounters. This research can have a high potential for bias due to the researcher's involvement in the process, for example, a female researcher observing village life in a tribal environment will have a very different experience than a male researcher. This method was not suitable here as it would have provided a view at a single construction site and any findings would not be sufficiently generalised. Equally, the outcomes would not necessarily contain personal insights and reflections from the workers, but rather observed behaviours.

- Survey.

  There are as many different formats of survey questions as there are written forms of communication. From this broad range, a few structured formats are used the majority of the time: ranking or Likert-type questions; multiple-choice or categorical; open ended questions with free-text; and dichotomous aka yes/no questions. However, questionnaires can also contain visual answers and unusual question formats. Online surveys can allow a researcher to gather data from a wide geographic area and from different demographics in a relatively short time frame. They also collect the data in a digital form which can speed up future data processing. However, this method does not allow engagement with the participants to prompt more detail or ask clarifying questions. Also, survey response rates are notoriously low, especially for free-text questions which would be required for this investigation.

- Document analysis.

  This would allow observation of the *formal* processes and attitudes in place to document failure. By '*formal*' I imply processes and attitudes approved by the organisation. This would give a biased, or one-dimensional, view of the 'truth' and would suit a multi-method approach where this version of truth is offset with other data. While the outcomes of this method would be interesting and beneficial, it would only achieve a partial picture and was discarded as there was not time to pursue multiple investigations. Future work could consider undertaking this method to contrast with the findings presented in this section.

- Focus group.

  A focus group allows engagement with multiple participants at once. However, I felt that the social dynamics in a such a group would encourage the 'acceptable' or 'expected' responses rather than the true reflections of all the individuals. Having said this, the additional insights gleaned from analysis of the discussion development over a focus group time period would have also been interesting, e.g. "do people gradually discuss failure with more candour? How does the researcher's role as a facilitator affect the openness of discussion? Do opinions change?" However, these questions were not the intent of this investigation and therefore this method was disqualified in this instance.

- Interviews.

  Interviews allow direct conversation with participants, and questioning can generally be split into three main types: structured, semi-structured and unstructured. Structured questioning is effectively a spoken survey whereby the questions are preset and each participant is asked the same question set. Conversely, truly unstructured questioning has no preset agenda or questions but is rather a conversation with an individual. These two methods would be too closed and too open respectively. Semi-structured interviews allow a fluid format to the discussions, including clarifying questions, meanwhile ensuring the relevant topic areas are covered (Harreveld et al., 2016). While the number of participants would be lower than that of a survey or series of focus groups, I felt this data collection method was most suitable and aligned with both the objectives and philosophical stance of this investigation.

I surmised that semi-structured interviews were the most suitable method of data collection. I also considered the data analysis methods and concluded that thematic analysis of the interviews would yield the desirable level of insight and understanding, allowing inductive analysis of the data.

As previously mentioned, in order to avoid unconscious bias in this area, especially confirmation bias, close examination of the literature was withheld until after completion of the interviews themselves. In hindsight, a useful exercise would have been to conduct a self-assessment prior to commencing any interviews, to explicitly capture my own values and belief on the subject. In this way, I could have used these to check for biases and add depth to the analysis. Unfortunately, this came to me in hindsight and, as such, I had to conjecture what my beliefs were at the time and judge whether any bias was present.

In total, 19 semi-structured interviews were conducted with members of the construction industry across several infrastructure sectors at different levels of business. The interviewees were approached through mutual professional acquaintances and Table 4.1 shows a demographic summary of the interviewees. By approaching the subjects via mutual acquaintance, the response rate of interviewees is increased and interview process is sped up. However, there is potential for bias as the interviewees are already within a subgroup of the construction population. Additionally, there may be an element of influence from the power dynamics from the mutual acquaintance to the interviewee, especial if the mutual acquaintance is in a position of authority. I took this into account by ensuring interviewees understood the confidential nature of the interviews and by empowering the interviewees to express their personal views. Power dynamics are further discussed in the next section.

Table 4.1: Interviewee Demographic

Interviewee Number	Time in Industry	Age	Gender	Sector	Title
1	26	45-55	Male	General Infrastructure	Technical Director of Infrastructure
2	28	45-55	Male	Rail	Programme Director
3	7	25-35	Male	Rail	Head of Programme Management
4	15	35-45	Male	Rail	Senior Project Engineer
5	30	45-55	Male	Rail	Senior Design Manager
6	20	35-45	Male	General Infrastructure	Group Learning Manager
7	20	45-55	Male	General Infrastructure	H&S Advisor
8	50	55+	Male	Renewables	H&S Manager
9	25	45-55	Male	General Infrastructure	Exec Corporate Development
10	27	45-55	Male	Rail	Programme Manager
11	31	45-55	Male	Rail	Quality and Reliability Manager
12	10	45-55	Female	General Infrastructure	Environmental and Sustainability Manager
13	3	<25	Male	Rail	Graduate Business Improvement Engineer
14	13	25-35	Male	Renewables	Site Manager
15	34	55+	Female	General Infrastructure	Commercial Services Director
16*	35	55+	Male	Renewables	Client Representative
17*	15	35-45	Male	Renewables	Client Representative
18	19	35-45	Male	Structural Design	Technical Director
19	25	45-55	Male	Structural Design	Commercial Director

*No transcript of interview

4.2.3 Interview Development

Semi-structured interviews do not contain a set list of questions which interviewees must answer but rather talking points which prompt discussion on the research topics. Prior to the interviews, I developed a set of question prompts which would initiate discussion. In designing the interview prompts, neutral language was aimed at to avoid bias or leading questions. Particular care was taken to ensure interviewees were empowered to provide their own opinions and not to report the 'expected' answer.

Despite every effort made to mitigate against potential bias in the interview data, there are still several ways in which bias may have crept in. For example, while limiting literature searching until after the interviews will have helped, the interviewers' preconception of what is and isn't important/relevant will have encouraged the conversation on certain routes of enquiry and possibly neglected others. Especially in later interviews, reoccurring themes may have featured more prominently due to unconscious confirmation bias. As a researcher, I was cognizant of this fact during the interviews, and later during the thematic analysis of the data.

The social dynamics of the interviews should also be taken into consideration. For a majority of the interviews, I was the sole interviewer and, as a junior female, would have been nonthreatening and (hopefully) approachable. The body language of interviewees and language use in their answers was noticeably different in a couple of interviews when Dr Simon Smith, my primary supervisor, was in the room. Additionally, a final two interviews were conducted by an MEng student, with myself supervising. The student had prior relationships with her interviewees having worked in their office during summer placement. This again changed the dynamic. I discuss this further during the narrative in the next section.

Empowering interviewees has been found to be essential in eliciting full and open conversation which reflects their personal beliefs and values. How an interview begins can set the dynamic for the rest of the sessions. During these interviews several steps were taken at the beginning of each session to empower the interviewee.

First, I used consent forms. These are not only essential for ethical reasons but also serve to empower interviewees by alerting them to their rights and providing them the opportunity to request to review their transcripts and/or any notes. By setting these provisions out at the start of the interview, the interviewee was put in a position of power over the interviewer.

Secondly, care was taken during the set up of the audio recording device to place it out of direct line of sight and to ignore it once the interview commenced. Again, this is to empower the interviewee as many people get nervous at the idea of being recorded.

Thirdly, the first set of questions were about demographics and career history. While these included important information to give a demographic overview of interviewees, such as age range, time in industry and occupation, this section of conversation also aimed to ease the interviewee into a conversation flow as most people find talking about their own factual history easy. The main body of the interview then followed.

The research questions for this sub-investigation, after rephrasing, were:

1. What does 'failure' mean to different members of the construction industry?
2. How is information about failures captured and what are the barriers?
3. What happens to these data?
4. How is learning from these mistakes/failures currently implemented? Do they work?
5. How prevalent is an '*error avoidance learning climate*'?

An additional research question (5) was formed during development of the interviews. 'Failure' is a sensitive subject for many and I wanted a way to explore that sensitivity which empowered the interviewees to divulge. Nikolova et al. (2014) had previously explored this topic via a Likert-type questionnaire and had defined an 'error avoidance learning climate', where employees feel anxious to admit or discuss mistakes. In developing the prompts for this research question, I adapted examples of their Likert-type questions. This is the question which has the most initial bias as it was based on my assumption that there would be a level of anxiety or wariness concerning discussing past failures, both with me as a researcher and in the general workplace.

The full set of conversation prompts used is:

1. Failure
   (a) What does 'failure' mean to you?
   (b) Can you think of other types of failure within the construction industry?
2. Information capture
   (a) How do you deal with this kind of failure?
   (b) What processes are in place to capture 'lessons learned'?
   (c) Have you been at any previous companies who approached this differently?
3. Lessons learnt
   (a) How is learning from these lessons currently implemented?
   (b) Are they used efficiently/as designed?
   (c) Do you feel more or something different could be done? What?
4. 'Error avoidance learning climate'
   (a) Do you feel you can discuss past mistakes with colleagues?
   (b) Do you feel employees are anxious to openly discuss work related problems? (Nikolova et al., 2014)
5. Is there anything you feel the industry could be doing to promote learning from past mistakes?

As seen, research question (1) was split into two questions. The first question was kept deliberately vague but personal in order to prompt a response which covered what the interviewee thought were the highest priority failure types and also to see if any strong attitudes manifested at this point. Meanwhile the second was formed as a more direct 'catch-all' to extract what failure types the interviewee believed existed in the industry - regardless of their importance.

Research questions (2) and (3) were covered by the same set of prompts as I deliberately left them open ended. This allowed the interviewee to describe what came first to mind - information capture or analysis - this in itself indicates the importance levels they assign to parts of the information cycle. Again, I included a question about previous companies to not only extract a more generalised picture but also reinforce to the interviewee that this was about their experience of learning from failure in their career, not a company-led assessment of the current practices. My hope was that this would further empower the interviewee, and enrich the data collected.

Prompts about learning were phrased so as to first capture formal ways of learning, and then extract opinions on interaction with these processes and opinions of less formal methods.

Finally, I included a final question asking their opinion about promoting learning from failure. This was aimed to be an empowering question which would provide insight into personal priorities as well as add context to the further research. In my experience when using this prompt, the conversation went one of two ways: nothing to add or started a large discussion which provided the most insights on learning in the construction industry.

These question prompts were not all used every time as occasionally the conversation flowed easily to cover the subjects without prompting. Additionally, the order of the conversation differed in each interview depending on the interviewee, their experience and how comfortable they were talking about each topic. While I found, in general, that the conversation tended to flow better if I asked about factual processes first and sought opinions later; once a rapport was established, a couple of the interviewees took the opportunity to have a 'rant' about the state of company culture, essentially skipping straight to the section on error avoidance learning climate. The semi-structured interview method allowed this easy flow of conversation and allowed best extraction of the raw information from interviewees.

4.2.4 Data extraction and analysis

Data were acquired from the interviews via thematic analysis, aided by NVivo software, of both interview notes and transcripts, which were typed verbatim but did not include indication of pauses and intonations. Thematic analysis is a standard method used by social scientists for qualitative research and is an iterative method used to draw out underlying themes (Silverman, 2013). When properly implemented, it can be powerful at identifying key factors within context, and correlations which aid the formation of hypotheses. It should be noted that analysis in this way cannot prove causality, which would be better shown in a more experimental or action research method. For the research questions posited here, thematic analysis is a suitable method of analysis.

Analysis was initially developed by examining the data for key pre-identifiable theme areas, such as failure type, and developed further as new themes emerged. NVivo software is a tool to digitise analogue data analysis methods, primarily for social science. In this instance, it acted as a large mindmap. Traditionally, themes were found using highlighters, bits of string and post-its. NVivo facilitates this in an easily accessible digital format, allowing more connections to be captured and many more themes or 'highlighter colours' to be used. NVivo allows sections of data to be assigned to multiple themes and facilitates additional manipulation, such as filtering and comparison, which shows co-occurrence of themes that can then examined for correlation and comparison to generate theories.

The first set of themes examined in this iterative process were any mentions relating to the information management following a failure of any type. These utterances would be highlighted and added both to the relevant information management node and failure type node. For example, the following quote, "*we start out with an observation card, which is filled out when we find something untoward in an unsafe condition. An unsafe piece of kit or an unsafe practice*", was coded to both the 'Health and Safety' theme and the 'Information capture' theme.

During the transcription and coding of these first themes, I became familiar with the interview data and began to identify further themes which emerged during these conversations. This second set of themes capture stimuli and opinions which affect the attitudes or behaviour of construction employees in regards to learning from failure. These 'softer' aspects of learning from failure are important for any further research on this subject, so that appropriate methods and implementation plans are undertaken.

4.3 Thematic Analysis of Interviews: Narrative on 'People and Data'

This section is divided into three subsections: failure mode identification, learning process identification and attitudes to failure.

The first explores the different failure modes identified by the interviewees.

The second gives a narrative on the formal and informal processes identified from the interviews which construction employees use to learn from failures. This is then compared to learning theory in order to inform the future method selection. It also ensures that this research is directly applicable into the current UK construction industry, thereby maximising impact.

The third subsection explores the attitudes towards failure implicated during the interview. During the thematic analysis of the interview data, it emerged that a few key stimuli were affecting the learning processes and employees' attitude towards them. Not only are these important results in their own right, but also have significant implications for the methodology and potential impact of this research.

4.3.1 Failure Mode Identification

All interviewees identified several project 'failure modes', where a 'failure mode' is a type of negative consequence of an event. For example, 'money' encompasses overspend and unanticipated spend. Figure 4.1 illustrates all the failure modes identified during coding the interview data. In this Treemap, each block area depends on the number of interviewees who mention a failure mode. Three core modes, consistently identified in discussion, were: H&S, time, and money. The next largest modes were quality and structural safety. These identified failure modes are all well documented consequences of risk in engineering project management (Munier, 2016).

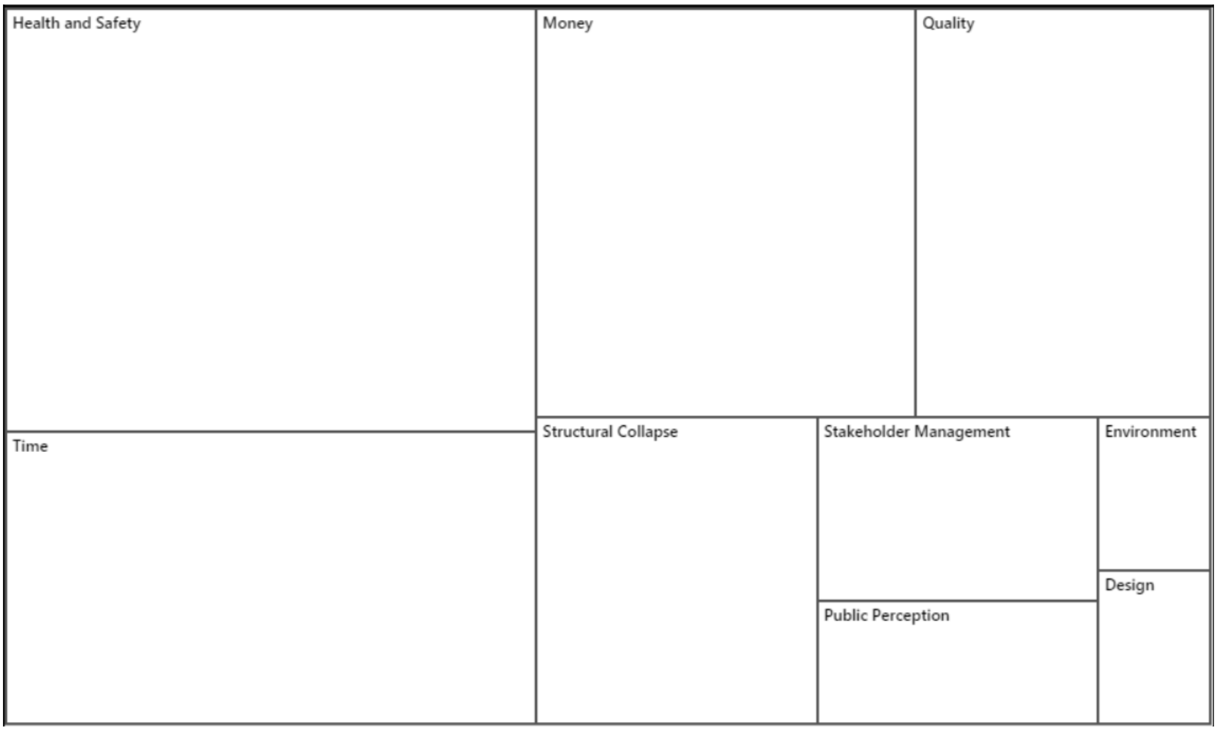


Figure 4.1: Failure Mode TreeMap

When comparing these ‘failure’ criteria to the ‘success’ criteria evident in previous literature (see Subsection 2.2), Health and Safety was far more prominent in the interviews than in the project success literature. This indicates that it is more relevant for management of failure than success. These distinctions are why it was so important to carry out this investigation - defining failure from the standpoint of investigating failure rather than implying failure from the inverse of success.

Several interviewees considered these failure modes through a root cause lens. For example, poor quality leads to expenditure, time spent on remedial and can result in an unsafe condition or safety incident. This created a hierarchy with some failure modes feeding into others. Figure 4.2 illustrates all the connections extracted from the interviews. Interviewees identified ‘time’, ‘money’ and ‘health and safety’ as the three key modes which lead to project failure. Other cited failure modes were deemed to be sub-categories of these as the project failure. As summarised by one interviewee: “*the others all feed into these three*”.

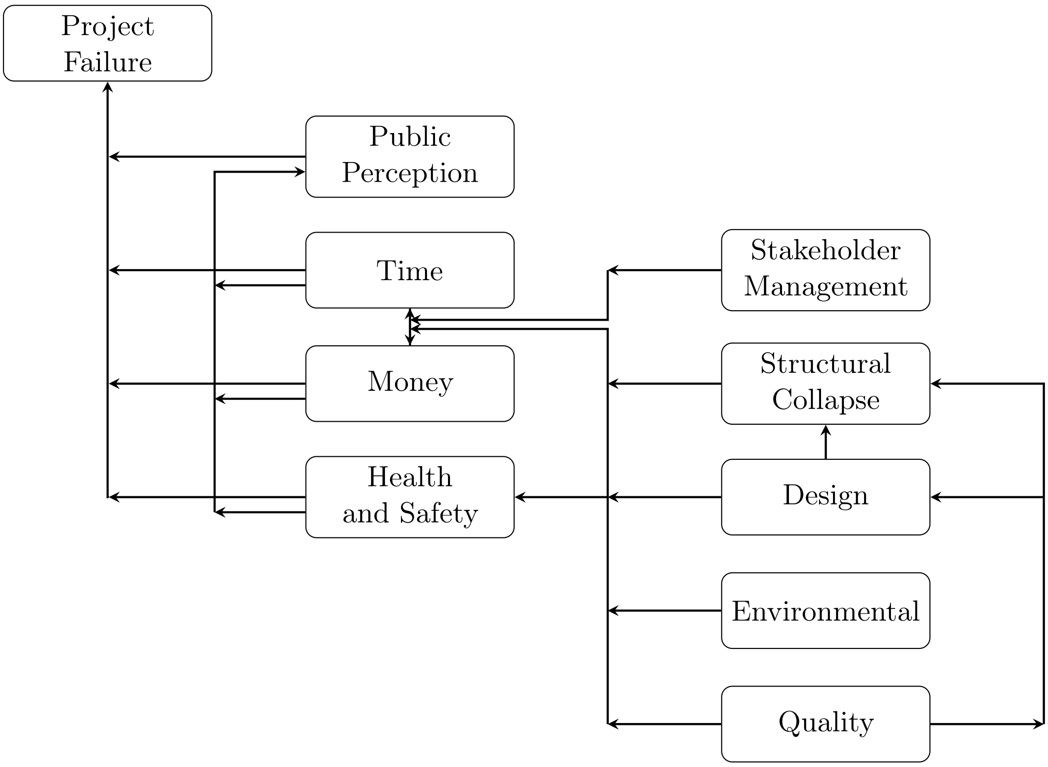


Figure 4.2: Failure Mode Connections

The only mode, identified by two separate interviewees, which does not directly feed to one of the ‘top’ three is ‘public perception’. In fact these ‘top’ three feed into this. One interviewee explained that a project could fail by bad public perception “*because something fails or because they don’t like it, it doesn’t do what it’s supposed to do.*” This clearly indicates that, while the other failure modes can contribute to a failure in public perception, it can also fail even if all these other modes are successful, simply because public opinion has turned against the project or the solution was incorrectly identified.

Interviewees also identified that ‘money’ and ‘time’ failures are often conflated on project. These two failure modes are tightly coupled and can be exchanged for the other in many cases. For example, Interviewee 1 explained that “*what tends to happen if a job overruns is they chuck resources at it*”. This illustrates the extremely interdependent nature of time vs money failures on project. This is indicated by the double arrow between the modes in Figure 4.2. This interdependence is also reflected in the success criteria literature.

Additionally, and not reflected in the figure, interviewees implied indirect correlations or feedback from 'time' and 'money' pressures to the other failures modes. Three interviewees (3, 5, and 10) directly referred to 'pressure', either time or money, having a negative effect on other performance. For example, Interviewee 3 stated: "*there's a very set amount of time when the public aren't using the railway and things like that, which can generally have a negative impact on that safety culture because there's a lot of pressure to get things done.*" This clearly implies that having a time pressure has a negative impact on H&S performance.

Interestingly, environmental failures, such as spills and disturbing protected habitats, were only identified by the Environmental and Sustainability Manger, no other interviewees picked it up as a source of failure. Despite construction's recent drive to appear more environmentally friendly, this is not evident in the priorities expressed by the staff.

4.3.2 Learning Process Identification

While analysing the responses about how learning was implemented from these different failure modes, it became clear that there were defined stages of learning from an individual failure. This single-loop learning cycle was characterised by an initial information gathering phase following an incident followed by a period of initial remedial action and alerts. Some of these incidents then progressed to a long-term change or formal learning implementation. This cycle matches the generic stepwise learning cycle set out by Drupsteen and Hasle (2014).

Additionally, while the different learning processes identified in this analysis were consistent across different companies and engineering specialities, the maturity of some aspects varied depending on sector. For example, Interviewee 2 noted that, working in rail, he expected engagement with reporting NCRs (non-compliance reports) to be less than the nuclear industry but ahead of general building construction.

Safety

Safety was the most mentioned failure, with all the interviewees except the two client representatives stating that it was a potential form of failure within the industry. Moreover, 12 of the 19 interviewees identified H&S failures, such as incidents involving injury, as the focal form of failure in the construction industry.

Of the identified failure modes, interviewees recognised learning from safety failures as mature in respect to the paperwork and formal process. One interviewee stated that:

> *Safety legislation is there, [. . .] I think for me dealing with safety and minimising failure, it's a state of mind and it's a culture*

This was reinforced by other interviewees who were pleased by the current formal system and referred to the process as industry standard, although several acknowledged that there were still steps to be made to improve the uptake and personal buy-in of certain learning stages. Additionally, there is a wide belief that more needs to be done to drive these processes down to contractors and SMEs.

Overall, the safety learning cycle was presented as a closed, well-standardised single-loop learning cycle where information is collected, analysed, distributed and then stored. Interviewees tended to be content with this learning cycle for larger incidents; however, felt that it was insufficient for smaller events as there was a weak link in the learning cycle which would fail. For example, the small incident was not recorded or it would prove too costly in terms of time and/or resources to investigate it.

Especially focusing on communication of incidents, one interviewee said:

> *I think health and safety is one of the few examples which broadcast outside companies and [...] they take the learning across the industry. Because it involves people's lives, you know, [...] it's people's actual lives that are at risk and so it's a bigger issue than the individual company.*

This indicates a deep level of commitment to this learning cycle, buy-in at all levels of the organisation - not just top-down or bottom-up.

The buzzword on people's lips seems to be behavioural science or developing a positive safety culture which was mentioned explicitly by 7/19 interviewees. The inclusion of values and culture into the learning cycle marks the migration from single-loop to double-loop learning. This type of learning could tackle underlying issues which are currently inhibiting learning. However, Bye, Rosness, and Røyrvik (2016) note that the attention given to culture could be a 'two-edged sword' as the use of 'poor safety culture' as a reason for incidents might lead to premature closure of an investigation into root causes which are key to efficiently reducing reoccurring failures (Haslam et al., 2005).

Safety Case Study: Thames Tideway

This case study presents my own experience on the Thames Tideway project which demonstrates the shift towards behaviour science and developing a positive safety culture. Thames Tideway is a multi-billion-pound project in London to upgrade the existing combined sewage system, which collects both sewage and rainwater. The preparatory works began in 2015 and it is due to be complete in 2024.

Before working on any Thames Tideway site, every member of staff is required to attend the EPIC (Employer – Personnel Induction Centre) Induction Day. This Health and Safety focused day was based around a live-action roleplay type example of a fatal incident. It explored the role of the company culture and communication, as well as the external pressures and further consequences of a serious incident. This fully immersive day was an extremely effective way of conveying the complexity of such situations and the ease with which something could escalate.

The impact of this day was made more powerful by the analysis coached from the participants, to show that intervention at any level of the people involved in the incident could have prevented the catastrophic outcome. This aimed to empower those on the course in their own daily roles, whether that be CEO or general worker. Additionally, the transition in the afternoon to demonstrative examples of what creates good and bad communication was a good practical step toward fostering a collaborative atmosphere.

Overall, the day espoused an idealised message of safety first and promoting healthy dialogue across the project, focusing on behavioural safety.[a] The details of this training day are included on the host company's website *EPIC* (2016).

[a] Unfortunately, I thought that this conflicted with the everyday experience on the project which undermined the purity of this message. For me at least, there was a consistent message of cost cutting which reduced the confidence I had in the prioritisation of safety over cost.

Quality

Non-compliance and poor build quality was identified by half the interviewees as a specific failure. While the initial learning process presented by interviewees is extremely similar to that in place for H&S, there were more concerns over under-reporting, lack of analysis and inadequate feedback. Several interviewees were keen to point out that there were systematic quality checks in place to avoid non-compliance reports (NCRs) including managerial reviews requisite under ISO 9001. Interviewee 10 stated

> *Generally quality is quite well-managed, we use quite tight process to ensure we use the correct products and the correct stuff and that it's all approved.*

However, this active management generally refers to managing quality prior to failures or implementing remedial action to ensure the quality of the end-product, not implementing systematic learning from failure. The majority of interviewees were pleased with the level of immediate response of an investigation and remedial action; however, they found that long-term trends and learning opportunity were lost into the blame game. The general message was that NCRs were used actively on projects for firefighting and remedial action; however, there was far less engagement with analysis than H&S. Interviewee 1, a technical director, stated that they probably do nothing with the reports, acknowledging that there should be some kind of statistical analysis to identify trends similar to H&S data.

Reporting engineering non-compliance (NCRs) was referred to as a "little bit scary" and it was indicated several times that people were more willing to put in snag or improvement reports as the personal consequences were seen as less severe. The exception to this rule was when the potential safety consequences were judged to be serious or life-threatening. Discussion of new technology for reporting presented an interesting conflicting view where a younger interviewee remarked that it made reporting quicker and easier to store, while an older interviewee stated that it made reporting more opaque and less accessible to those on site.

In comparison to safety, therefore, quality had a far less complete single-loop learning cycle as, while information is captured, very little analysis and extremely sparse distribution occurs. Equally, while the information is generally electronically stored, this tends to be silo-ed by project, rather than in a central data repository, and access is limited both by permissions and opaque search tools. Nevertheless, it should be noted that interviewees gave good examples of informal feedback and team discussion to analyse or learn from serious examples of these events. These unformatted lessons learnt exercises were occasionally captured for future learning but interviewees were very sceptical as to their worth.

Time and Money

Time and money were also identified as key factors in defining project failure; however, learning from incidents of overrun or exceeding budget were less well defined and varied greatly between levels of the business. These failure modes refer to more commercially sensitive root causes and are not as easily captured.

Tacit learning was, therefore, the only identified method of on-job learning along with some mention of generic formal training courses. Consequently, innovations within this section of business are kept within a very small community. Executive groups or small communities tend to share their internal learning using discussion such as informal lessons learnt sessions. Interviewees working in these areas did not feel it inhibited their individual learning on project as the teams are small; however, they acknowledged that staff turnover and lack of formal capture restricted learning outside each project.

While accounting records and schedules should record changes and why these events occurred, there is no systematic cyclic assessment and feedback/distribution of information within (or outside) the business. Although 'notice to delay' exists, its use is misconstrued and therefore not used properly. The lack of systems approach for cost overrun has been explored by Ahiaga-Dagbui et al. (2016), however no robust methods have been suggested for improving capture and analysis of this failure type.

Other failure types

The two interviewees who mentioned public perception as a possible failure mode did not identify any associated organisational learning cycle with this failure. Meanwhile, stakeholder management, identified in three interviews, was described as a human process, with the management strategy focused on personal relationships. Unlike other personal communication, interviewees described this activity as a "*rigorous process*" and managed by an "*effective strategy*" suggesting that this form of communication is not left to grow organically, but is at least guided by an underlying ethos. However, there was no formal learning from this failure type, with the process focused on reactive action to failure. One interviewee described this process:

> *[..] go and see them or talk to them [to] try and see if there's any way can resolve or lessen the impact of the failure in the relationship. [...] It's about making sure they feel listened to and that they're being taken seriously and that somebody's going to do something about it.*

Structural collapse was mentioned by five interviewees. However, they indicated that this failure mode is strongly related to safety or quality failure, suggesting that the collapse itself could be considered the immediate cause of the failure rather than the failure criteria itself. Quotes to support this include:

> *The whole building collapsed because they made several different quality mistakes*
>
> *A structure or temporary works may have collapsed, killed somebody or injured somebody.*

In these cases, the collapse is immediately related to another failure criteria which is captured in its own process. The conflation of safety and quality in this case remains a theme throughout these five interviews: incorrect quality can cause a collapse which is an unsafe condition. Therefore, should this failure be captured as a safety observation or a quality issue? In reality, interviewees noted that before a collapse occurs and unless a safety incident actually occurs, these faults could be picked up as quality or design issues. However, once a collapse reaches a certain severity, this failure can trigger a forensic investigation, such as described in the Edinburgh School's case study in Section 2.3. This type of failure is then analysed in detail and lessons are communicated around the industry; however, in practice, one interviewee noted that "*unless it's a catastrophic thing, they won't necessarily pass that on as best practice*".

With the exception of the two designers, Interviewees 18 & 19, the interviewees were based in the construction delivery stage. Design as a failure mode was only identified by one construction delivery employee. This could be because design issues on site manifest themselves as different failures. For example, as a quality issue - not compliant, a safety issue - design was unsafe to construct, or time/money issue. Therefore, the failure was collected as one of these types. The designers interviewed noted that they and their team learnt from design failures through team discussion and mentoring. This learning was focused on personal competence, rather than organisational processes. It should be noted that only one design company was interviewed and therefore these findings cannot be generalised to other designers.

The Environmental and Sustainability manager indicated that environmental failures, such as spills and disturbing protected habitats, are recorded in a similar manner to safety and quality failures. Data are captured after an incident, and the information used on site for remedial action and to implement measures preventing re-occurrence. However, similar to quality, the learning from this is an extremely incomplete single-loop cycle with very little analysis, extremely sparse distribution occurs and data silo-ed by project/site. Unlike quality, there is an external body - the Environment Agency (EA) - which the projects have to answer to if the incident is significant. It does not appear, however, that this agency is being included in the learning cycle in the same manner that the HSE is helping to facilitate learning for safety failures.

Another learning process identified by several interviewees is formal "lessons learnt" documents at the conclusion of the project, or milestone in the project. This is not focused on any one failure mode, but aims to capture learning which project participants discovered during the course of the project, often by running a workshop-style session, and record these lessons for others to refer to. However, interviewees felt that these exercises failed to achieve their aims for two main reasons: (1) the key people involved in the project have moved on or weren't present; and (2) limited consumption due to time pressures. The first reason is a barrier to data collection, while the second is a barrier to knowledge sharing. One interviewee, new to his project, expressed his frustration with trying to use lessons learnt documents for that project as he tried to understand what had gone wrong in previous stages: "*for the lessons learned, they didn't necessarily invite all the right people or everybody to that forum*". He describes how the information is biased and fails to be as useful as hoped, despite taking the time to read it. The time investment required in finding and comprehending these documents also limits their usefulness. Another interviewee describes how he sees the issues with the current database style access:

> *Lessons learnt databases are massive and voluminous and people have the day job of running the other projects. You don't have time to lock yourself away in a room for days and look through to understand the issues that have happened in other jobs*

4.3.3 Attitude to Failure

"An attitude is a tendency to respond in a consistently favourable or unfavourable manner towards a specific topic, concept or object" (Miller and Brewer, 2003). While failure as a whole could be taken as the concept here, there are several separate issues that stem from failure which were found to drive certain behavioural responses. These are subsequently referred to as attitude stimuli.

During analysis of the interview data, key attitude stimuli were identified with their corresponding responses. Two pairs of these stimuli will be discussed here: blame and ownership; leadership and acceptance.

Ownership and Blame

A theme which emerged was reluctance to take ownership of the failure. Multiple interviewees alluded to this with a few citing reasons such as: not good for your CV, if I knew my job wasn't on the line and it's very painful, it's embarrassing. One interviewee pointed out that directly employed members of staff or those employed by the main contractor were more likely to raise an issue as he put it they feel ownership because they are part of a larger group. There was also mention that by specifically referring to job security and the length of work during inductions, the site workers tended to be more involved in the job, rather than just carrying out the assigned task. This concurs with recent emphasis in research, such as (Sanne, 2008), on increasing employee ownership to cultivate a productive reporting procedure.

On the other hand, for failures where there existed an overwhelming sense of moral obligation to take ownership, interviewees expressed increased satisfaction at the learning process. For example, H&S failures have a moral imperative to help preserve life and quality of life to others.

This was expressed by one interviewee succinctly:

> *Everyone is very open-minded about sharing lessons learnt from safety incidents because of the overarching moral obligations*

Perhaps due to the different amount of perceived moral obligation, different failure modes seemed to elicit different levels of personal or company ownership. In comparison to H&S as already outlined, discussion on quality failures led more to blame and legal consequences, for example contractual conflicts. Additionally, if quality processes can be improved by a certain action, it is in the interest of the company to keep it undisclosed as a Unique Selling Point. Such reasoning overlooks the interdependent nature of quality and safety in construction where investigations have indicated mutual causality, where each performance type positively impacts the other (Wanberg et al., 2013; Love et al., 2015). Given this, the industry should ask itself: is it morally justified to keep back significant quality information?

Reluctance to take ownership had significant co-occurrence with the theme of personal blame or consequences. Some of the many quotes on the subject were:

> *We live in a world of blame culture. Whether you like it or not. People always worried about being the one at fault. You got your battle lines drawn very quickly.*

This discourse of blame and fault is at odds with recent research and policy to foster a no-blame culture, especially within H&S, to not only address learning but also encourage collaboration and innovation, for example (Lloyd-walker et al., 2014).

An interesting finding was the role interviewees perceived HSE to take in regards to H&S learning within industry. Several times, it was hinted that inclusion of an independent body within the learning cycle shifted the internal focus from blame and personal culpability to learning and fair distribution of information. The legal obligations also gave professionals within the H&S industry an external scapegoat to avoid internal conflict as Interviewee 7, a H&S advisor, noted he was able to say to site staff in relation to enforcing H&S that "it's not just me once or twice a month, HSE could come up here any time".

Acceptance and Leadership

Acceptance of failure, or rather the lack of acceptance, emerged as an important attitude stimulus within the discussions with interviewees.

> *They go: [. . .] "It will never happened to me."*
>
> *People [. . .] think "oh, we'd never do that on our project."*
>
> *I wouldn't say we had any failures.*

This topic co-occurred with discussion of the role of leadership and top-down incentives for encouraging learning from failure. It was explicitly stated that increasing incentives and the acceptance of failure will aid prevention of failure:

> *I think people should be incentivised to produce these things and to accept the fact that we've got something wrong. Because, if you don't accept the fact that you've got something wrong, you're never going to prevent those things happening.*

It was indicated by several interviewees that learning from failure is lacking incentives. Several interviewees noted that leadership are often given financial incentives for productivity or profit which is in direct conflict with the acceptance of failure. Also on a personal level, one interviewee notes that a project which was considered a failure is bad on your job record. However, projects are an amalgamation of the work and effort of a (sometimes huge) number of people and the overall success or failure of a project rarely reflects on the specific value you brought to the job or the valuable learning gained from this. This observation can also be scaled up to the company as, when bidding for work, successes are emphasised, and failures unheeded. One interviewee explained the situation nicely:

> *When you tender for work, clients will ask you what you got right, never ask you what you got wrong and what you learn from it. [...] I find that's an interesting way of just ignoring it basically. You don't get repeat business by broadcasting failure.*

In fact, this limited openness was evident in the interviewees when the power dynamic of the interviews were changed. As previously mentioned, the majority of the interviews were conducted by myself - with no others present. However, a minority of interviews were conducted with Dr Simon Smith present. Although not participating in the interviews themselves, it was clear that having a more senior, male person overseeing affected the confidence of the interviewees. They used more neutral/careful phrasing; however, I noted that this did not necessarily result in less candour in the content of the conversation in every case.

The most closed interviews occurred during the interview of Interviewees 16 & 17. These individuals refused audio recording and insisted on a joint interview. By having another participant present, they effectively censored each other. This interview was the only one where I felt the interviewees were deliberately sanitising their responses. This could be due to having a peer present who could judge their response, therefore they felt like they had to stick to the expected dogma.

Additionally, a final two interviews were conducted by an MEng student, with myself supervising. the student had prior relationships with her interviewees having worked in their office during summer placement. It appeared to me that, while the prior relationship meant that the interviewees were at ease during the conversation, there was also a slight element of mentorship to their answer, where they were providing the information you would teach a mentee, not necessarily what you experience in reality.

It is not considered that the effects of these different power dynamics had a significant effect on the final results, with the exception of the interview with Interviewees 16 & 17. However, it is a clear indication that, while construction professionals can be candid about failure in an artificial environment and profess to want to engage in learning from failure processes, the culture and expectations of their peers and the industry can have significant effect on their actions.

4.4 Insights for Learning from Failure Data

Summary of thematic analysis findings

Having identified the perceived failure modes in construction projects, this qualitative investigation explored the different systematic learning processes undertaken in the construction industry and the attitudes towards learning from failure.

This investigation concurred with previous research on learning in the construction industry which emphasised the reliance on building human competency. One interviewee nicely summarised this:

> *Basically what generally happens is that people get experience from other projects, by assembling project teams you get hopefully a good mix of people with experience across the project with different aspects. They bring, with the knowledge, lessons learned which can be acted on at the project.*

For learning from failure processes, analysis of the interview data showed different stages of maturity in the learning cycle applied to different failure modes within a construction project. While safety showed mature single-loop systematic learning and some migration towards double-loop thinking, quality presented an undeveloped single-loop process. Time and money failures gave no indication of any systematic learning process; however, there was strong evidence of informal learning and discussion.

Given these different stages of maturity, development of learning from failure in the construction industry cannot be tackled by a singular approach, but rather by developing different aspects of the process for each failure mode. This investigation found similarities in the data collection and learning cycles for safety, quality and environmental failures, and therefore has grouped these failures and will focus on development on learning from failure processes from data of this type.

Within discussion of attitude to failure, two pairs of attitude stimuli were discussed: Ownership and Blame; Acceptance and Leadership.

Discussion on ownership and blame highlighted three outcomes:

1. Blame suppresses learning;
2. Increased ownership of failure cultivates a learning environment;
3. Inclusion of an independent organisation within cycle aids failure analysis and wider distribution i.e. HSE for safety failure.

Meanwhile, dialogue on acceptance and leadership revealed the need for introducing incentives for learning from failure and emphasised the impact of individual and company leadership on acceptance of failure as a possible concept.

This preliminary investigation supports previous research indicating that organisational climate plays a large role in the success of learning processes (see Section 2.3). It has also uncovered some key information about current learning from failure processes in the construction industry, including how failure data is currently collected and used. This is essential for the progress of this research.

What does this mean for industry?

There are several findings from this initial investigation which are immediately applicable for the construction industry. These are framed in terms of learning from explicit and tacit knowledge. As previously discussed, both of these types of knowledge are produced by failure events.

This investigation found that the construction industry relies heavily on the tacit failure knowledge stored inside the minds of individuals. Tacit knowledge is contexualised and evokes the types of emotions associated with failure, as expressed by interviewees, including embarrassment, anger or denial. For this reason, while individual events are powerful and can result in lasting learning for those involved, the transfer of this knowledge between individuals is difficult.

Currently, the construction industry appears to rely on the passive occurrence of socialisation, as defined by Nonaka (1991), to transfer this knowledge. In other words, several interviewees indicated that this socialisation mostly occurs spontaneously or part of natural conversation. However, reliance on spontaneous conversation for organisational learning is inefficient, especially considering that people don't often voluntarily share examples which embarrass them. Interviewees indicated that the exception to this general rule is on safety matters, where the moral obligation to protect their colleagues prompts experienced personnel to share 'words of caution' accompanied with anecdotes to strengthen their point. This apprentice-type learning is essential to organisational learning in construction. However, one interviewee did note that this type of information transfer can prevent innovation and lead to a 'this is how we've always done this' attitude.

The only organisational process which interviewees mentioned that formally facilitated deliberate transfer of tacit knowledge is during "lessons learned" exercises. During the associated workshop and compilation of the lessons learned document, participants actively reflect on and communicate about their own experiences. However, these processes currently focus on 'codification' - the translation of tacit information into explicit information - which can be accessed via "lessons learned" databases, rather than the social aspects. Additionally, as previously mentioned, these documents are written some time beyond the events themselves and often do not involve the correct people. Within these documents, there is also risk of bias - both by what people deem important and also positive bias. One interviewee stated that the lessons learned documents are often written through a 'rose coloured lens' where participants focus on what went right. This concurs with the previously discussed point about acceptance of failure and the necessity for people to have participated in positive/successful projects for their career/CV.

There are two ways which this research suggests the construction industry can immediately look to improve tacit learning from failure. These are: develop more formal socialisation processes and to rethink lessons learned processes. In consideration of both these suggestions, incentives to participate should be carefully considered and aligned with project and industry objectives.

For the first, mentoring schemes have been shown to encourage learning. The construction industry has relied on mentor-mentee type learning, such as apprenticeship programs, for years. In the UK, there has been a concerted effort to reestablish apprenticeship programs, such as the Modern Apprenticeship scheme, to address the shortfall in skilled labour since the end of the last century (Hogarth and Gambin, 2014). However, these schemes are 'front-loaded' (Winch and Clarke, 2010), focusing on the initial provision of essential and technical skills to pass specific criteria. There has also been criticism of their efficacy, with participant non-completion and cost to organisation working as disincentives to organisation participation Hogarth and Gambin (2014) and Daniel et al. (2020). In contrast, mentor relationships can be established at any level of career, generally focusing on career steering and 'soft-skills'. Chan and Moehler (2007) found that these relationships help to fill the detachment between the formal training and employers' skill requirements.

Informal mentoring is already present in the construction industry, with senior members guiding junior members, with limited mentor training or training agendas. However, this research suggests that formally supporting and developing these relationships could be beneficial to the lifelong development of skills and transfer of tacit learning in construction. Hoffmeister et al. (2011) found that the top 5 traits for mentors in the construction industry are: (1) good listener; (2) willing to share negative information; (3) comfortable around superiors; (4) allows an apprentice to make a mistake; and (5) willing to give negative feedback. Three of these traits directly refer to dealing with failure, highlighting the importance of learning from failure to this process. Mentors should be provided training in mentoring and guidance/structure on addressing sensitive topics. Specific to the construction industry, a barrier to this type of scheme is the nomadic nature of the contracting workforce. The constant movement of people means that these relationships will be hard to create and foster. Generally used for white collar employees, long-distance mentoring, using digital communication technology, could be considered. ECITB (Engineering Construction Industry Training Board) has just launched (in 2020) an e-mentoring programme for project managers in construction, involving a limited set of 6 mentors and 6 mentees, expanding their previous Oil & Gas scheme. Investigation into establishing this type of life-long tacit training is recommended.

Another way this research suggests the construction industry could systematise the tacit information transfer is to develop their lessons learned exercises. This research suggests that the primary aim of these exercises should be to maximise the tacit transfer of failure (and success) information, with capture of that information as a secondary priority. While the industry could also look to implement better methods of codification and retrieval of the lessons learned documents, using AI and smart search algorithms as in Eken et al. (2020), this does not address the root cause of why these databases are not used - people simply do not have time to use them.

In relation to explicit information, this investigation identified lessons learned exercises and failure reporting as primary forms of codification of failure data. While other data relating to failures exists, for example project managers reports or comparison of schedule data before/after delays, failure reporting systems exclusively document failure events. Failure reporting - safety, quality and environmental - creates reports which document the individual events. These reports are used re-actively on the project site for corrective actions and for feed-forward to upper management in the form of 'dashboarding', i.e. graphically displaying summarised data, and tracking performance measures. On one project I worked at the project team actively ensured feedback for these reports, discussing any from the previous day at the start-of-work brief, however, in general, the benefits of these reporting procedures were not seen by personnel on site.

Furthermore, much of the pertinent information in these reports is contained within the text description of the reports. Analysis of these text descriptions is currently restricted to manual techniques, which are resource intensive, inefficient and subjective. As such, this analysis is rarely carried out for the low consequence events. As hinted at in Section 2.5, modern AI and natural language processing methods could facilitate systematic analysis of these reports. This would allow the construction industry to better exploit these data to their full potential and implement the findings into organisational learning practices.

What does this mean for this research?

This investigation has demonstrated the need for a method to facilitate systematic learning from failure reports on construction projects.

Previous research has investigated NLP and AI to increase the effective analysis, transfer and access to the information contained within construction text. These methods have the potential to unlock the potential of these reports. This leads to the development of research questions (RQs) 3&4: first investigating the methods available to analyse the text-based failure data (RQ3), and then proposing how this could be incorporated into organisational learning in the construction industry (RQ4). Additionally, the outcomes of this initial investigation have several significant implications for RQs 3&4.

When identifying which NLP + ML pipeline would best facilitate knowledge discovery from text-based failure data (RQ3), it is clear that the evocative nature of failure on construction sites and the bias which exists in the data requires a human-centred approach. 'Human-centred machine learning' not only considers the human context in which the solution is being developed but also explicitly considers the human work involved in feature and model selection, gathering data, and decisions made by the analyst. Gillies et al. (2016) proposed "human-centred machine learning" as a phrase that "articulates a core set of values and approaches" in order to incorporate this thinking formally into ML development. While there are several relevant human aspects which emerged from the interview data, the most prominent in term of affecting the data collected was blame. Any system should aim to mitigate 'blame culture' which indicates a significant need for transparency in the ML and overall learning process.

This links to the concept of "explainable machine learning". Explainable machine learning covers an area of research which explores the ease and ways to explain model behaviour to various stakeholders. Bhatt et al. (2020) found that there were four needs to develop explainable models: debugging, monitoring, audit and transparency. They state that "Organizations that deploy models to make decisions that directly affect end users seek explanations for model predictions." As the aim of organisational learning is to instigate change, directly affecting those within the organisation, this is particularly relevant here.

These two value sets - human-centred and explainable ML - will be key in selecting and developing an NLP + ML pipeline suitable for learning from failure.

In investigating how this type of learning would be implemented into a systematic process for the construction industry (RQ4), a key factor identified is the lack of incentive or motivation to participate in such a process. While every interviewee espoused the need for learning from failure processes in construction, they also noted the lack of time and incentives to actually participate in such processes. Synchronising feed-forward and feed-back learning will be essential in ensuring these processes are effective. It is also essential to ensure that learning processes integrate with existing processes, reducing the additional time requirement and mitigating against becoming a 'box-ticking' exercise.

Limitations

While this investigation is extremely valuable to progressing this research, it is only a preliminary study into the the concept of failure on construction projects. The study also has a couple of main limitations:

1. Limited scope. The findings presented here were based on only 19 interviewees, who were all contacted via mutual acquaintances. This low number restricts the generalisability of the results - although it should be noted that no new failure modes were being identified from new interviews which suggests that thematic saturation had been achieved. However, the maturity and form of learning may differ for different companies and geographies. Additionally, a majority of the interviewees were employed in the construction delivery phase of the project. This is appropriate for this research project which focused on the project delivery phase of construction, however, to gain a better overall view of failure in construction projects, a more diverse stakeholder range should be considered.
2. Data collected. All data were acquired via semi-structured interviews. this method of data collection has inherent limitations. The limited time frame and formal style of data collection will have limited the information gained from the interviewees. Also, although mitigation was in place to empower the interviewees to give honest opinions, there will have been an element of bias towards telling the interviewer what they wanted to hear - i.e. "which failures do I think I should be identifying and prioritising?" rather than "which failure do I actually prioritise?". Some of these limitations could be addressed by supplementing this data with other data sources - such as project documentation and longitudinal ethnographic study.

> During the course of this research, I have also been seconded onto construction projects for periods of time as a site engineer. My own personal experiences during these placements will supplement the findings presented here. However, it will be made clear when this is the case.

Future research should aim to develop these findings, bringing in multiple data sources to improve the depth and generalisability of the results. A study similar to those comparing the success criteria/failures (as in Table 2.1 in Section 2.2) could be beneficial, especially as it would allow comparison between failure and success criteria - testing the hypothesis that the same criteria and factors which are important for success are also important to avoid failure. However, any investigation of this type should account for the previously mentioned limitation whereby previous authors of these surveys have not separated out different success criteria and so have aggregated all types of project and product success under one umbrella. As such these surveys also hold bias as to the success priorities of the professionals themselves. If failure criteria and factors were surveyed in a similar manner, the same biases would hold true.

Additionally, the messages delivered here can help focus future work on developing specific methods for learning from failure in construction that address the individual barriers identified by interviewees and wider literature.

Chapter 5

Unstructured to structured data: Using NLP & ML methods to structure unstructured failure text data

Structure *struc · · ture* | *strʌk − tʃə*

noun :

1. the arrangement of and relations between the parts or elements of something complex
2. a building or other object constructed from several parts

verb :

1. construct or arrange according to a plan
2. give a pattern or organization to.

Structure from https://languages.oup.com/google-dictionary-en/

The central concept for this chapter is ‘structure’. I find it appropriate that this word both expresses the end goal of a construction project as well as alluding to the solution to accessing information within the complex, unstructured data of construction projects. Identifying the key elements of these complex situations, and then investigating their relations, is required to exploit their learning potential.

5.1 Methodology and Method Overview

Methodology

Chapter 3 outlined the importance of establishing a methodological understanding of the research. In designing this research, I adopted a pragmatic approach in-line with the pre-generation of the problem statement. This led to the selection of post-positivism as the most suitable epistemology for this section of the research, as seen in Section 3.2.

Additionally, the qualitative investigation in Chapter 4 highlighted the need to incorporate human-centred machine learning values into the method selection and development of this research. Therefore, throughout this investigation, effort has been made to explicitly delve into the reasoning and internal logic behind why I chose to pursue certain routes and algorithmic methods and not others. By explicitly highlighting these decision points, the positivist epistemology is supported for this portion of the research. Luft and Shields (2014) explored the concept of incorporating subjective decisions in relation to the positivist nature of accounting research and succinctly described the logic: "*The positivist ideal of objectivity also includes explicit awareness and reporting of the subjective judgments and decisions involved in developing causal explanations and making research-design choices. [...] Reporting these limitations—that is, reporting the (often unavoidably) subjective nature of developing and validating causal explanations—can, perhaps paradoxically, increase the objectivity of a research study.*" As such, and in-keeping with prior writing in this thesis, the first person voice is used when subjective decisions are documented to highlight the researcher's role here.

Method

This chapter of research addresses the main part of the primary research question, RQ3: "Which Natural Language Processing (NLP) + Machine Learning (ML) model best facilitates lesson identification from text-based failure data?". Specifically, documented is the investigation of NLP + ML pipelines, which can be used to structure the unstructured text failure data and the results of the chosen process.

A key requirement of this data analysis, found by the qualitative investigation in Chapter 4, is to embed transparency and explainability into the AI method. Attribute-based analysis, developed originally for construction safety data, provides a clear, intuitive method which can increase trust. Therefore, the first step in the proposed method is converting the unformatted text data into a structured form by identifying fundamental attributes from the data. Further discussed in Section 5.3, research at the University of Colorado has previously developed an exemplar set of safety event attributes seen in Table 5.1 (Desvignes, 2014). This important step forms the majority of the research presented here. This can also be considered an intermediate step between data and information in the hierarchy of knowledge seen previously in Section 2.7.

The aim of this task is to produce a set of fundamental attributes which represent the text description in order to facilitate further analysis. An analogy for this process is the act of representing a landscape as a map. An aerial photo represents a huge amount of the information in the landscape, however, is too detailed and complex to easily interpret for the purpose of planning a journey. On the other hand, a line sketch containing the key features - stream/road routes and locations of buildings - is an abstract representation of the landscape, containing less of the information but is more useful to the journey-maker. In this way, refining the unstructured information contained within the text descriptions into a set of attributes reduces the complexity of the representation and allows analysis and interpretation. This analogy is returned to during the discussion.

Automatic extraction of these attributes from text data is required to make this a viable option for the construction industry. Manual labelling is too time-consuming.

Section 5.3 outlines the method decisions, based on previous work, and required theory. Section 5.3.2 describes existing literature exploring NLP + ML pipelines to analyse text-based safety incident data. This enabled an informed decision on the specific method task for this extraction. From examination of this theory, text classification is selected as the best method, over NER and predictive region detection. Prior to this, key ML concepts are defined in sub-section 5.3.1.

The extraction of work attributes by text classification is formed of two parts: development of the attributes through systematic labelling of the safety report data set, followed by application of NLP and ML to predict attributes in new safety event descriptions. This two-step approach is adapted from protocols developed and observed at the University of Colorado, Boulder (for example Tixier et al. (2016b)).

Development of this attribute list by manual labelling of text descriptions is explained in Subsection 5.4.1, while the resultant attributes are presented in Subsection 5.5.1. Section 5.4.2 documents the NLP + ML methods adopted for prediction of attributes in new text description data, while 5.5.2 presents the performance results these methods.

These attributes can then be used in many types of analyses to find patterns or trends in this structured data. To complete RQ3, the structured results from this chapter are then used in Chapter 6 to illustrate three use cases of knowledge discovery methods which can be used for learning.

Finally, in Chapter 7, these results are interpreted for learning in construction using the context discovered in Chapter 4.

Table 5.1: Examples of safety event attributes (Desvignes, 2014)

Upstream	Transitional	Downstream
Materials	***Equipment and tools***	***Human behaviour***
Cable	Forklift	Fatigue
Cable Tray	Hammer	Improper body position
Concrete	Hose	Improper security of materials
Concrete (liquid)	Ladder	Improper security of tools
Conduit	Light vehicle	Repetitive motion
Door	Powered hand tool	***Site charateristics***
Dunnage	Unpowered hand tool	No/improper PPE
Grout	Wrench	Object of floor
Heavy Material	***Materials and substances***	Poor housekeeping
Lumber	Bolt	Uneven surface
Metal studs	Electricity	
Pontoon	Hand size pieces	
Rebar	Hazardous substance	
Scaffold	Nail	
Soffit	Screw	
Spool Tank	Sharp edge	
Stairs	Slag/spark	
Steel sections	Splinter, sliver	
Wire	Small particle	
Equipment	***Site quality***	
Chipping	Cleaning	
Crane	Slippery surface	
Drilling	Unstable support/surface	
Formwork	***Weather and environment***	
Grinding	Mud	
Guardrail/Handrail	Poor visibility	
Heat source	Snow/Ice	
Heavy vehicle	Wind	
Job trailer	***Other***	
Machinery	Exiting, transitioning	
Manlift	Insect	
Stripping	Lifting/pulling/ manual handling	
Unpowered transporter		
Welding		
Design		
Confined workspace		
Congested workspace		
Object at height		
Object at height on same story		
Working at height		
Working below elevated workspace		
Working overhead		

5.2 Data collection

Construction projects collect and use failure data in many different ways, as seen in the qualitative investigation presented in Chapter 4. One of the key insights from this previous investigation was that any system to facilitate learning from failure should not create the need to manually collect more data, rather integrate into exiting systems. Therefore, data types which already exist in industry should be considered; although, this research will discuss and present recommendations to amend the collection and storage of this data to facilitate better learning value.

Another insight from the qualitative investigation concerns the pool of data sources for failure data. This initial investigation found that safety, environment and quality are the most documented failure criteria and that the data is collected and stored in a similar format. These data consist of reports on failure events which capture descriptive details of the individual events. They are generally captured via a form, either physically or digitally, and contain several different types of data. One set of fields collects structured data, identifying specific details such as date, time, location. These are directly comparable across the different failure criteria (safety, quality, environment). The second set of fields collect categorical data. These multiple choice type fields vary between failure criteria and collect information categorising the failure from a set list. The final field set consists of text descriptions. This can be a single field asking for an description of the incident, or several fields splitting the description or asking for additional information.

As identified in Chapter 4, these reports are used re-actively on site for corrective action and for recording the incident for posterity. The structured and categorical fields may also be used by management for limited trend analysis, however, the text data are currently not used in this manner.

Previous literature demonstrates analysis of text-based safety reports using natural language processing (NLP). Additionally, safety data are more accessible than quality data. The qualitative investigation indicated that this could be due both to the moral aspects of protecting people and legal responsibilities for reporting. This difference in accessibility was shown during the data collection, as the sponsor company stored all safety reports centrally in a common data environment (CDE) while quality (NCRs) were stored locally on project-specific environments. For these reasons, it was chosen to use safety data as a primary data source. However, unlike previous work which has focused on incidents which result in an injury, this research also includes safety observation reports. Safety observation reports document near miss incidents and unsafe actions/conditions observed on site. There are additional biases associated with this data which are discussed further in Chapter 7.

The data used in this research were gathered from a large UK infrastructure construction company. The primary data set consists of 14,266 safety incident and observation reports from a central Health, Safety and Environment database, recorded over a 9 year period.

These reports had the following fields which were filled out in y% of the reports:

- Structured/Numerical Data
  - Address (99%)
  - Postcode (89%)
  - Incident Number (100%)
  - Date & Time of Incident (100%)
  - Lost Days (16%)
  - Lost Hours (17%)
  - Postcode of Incident (25%)
- Categorical Data

  - Project (100%)
  - Framework (100%)
  - Sector (100%)
  - Division (100%)
  - Incident Category Type (100%)
  - Incident Type (100%)
  - Incident Sub Category (100%)
  - Location of Incident (100%)
  - Weather (55%)
  - Root Cause (19%)
  - Body Parts Injured (12%)
  - Injury Sustained (11%)
  - Overall Incident Status (100%)
- Text Data
  - Incident/Injury Details (100%)
  - Additional Incident/Injury Details (3%)
  - Root cause details (9%)
  - Activity being undertaken (30%)
  - Immediate Action Taken (78%)
  - Details of Action Taken (for individual action created) (6%)
  - Investigation Summary Comments/Conclusions (0.4%)
  - Suggested Improvements to Senior Management/Rest of Business (Lessons Learnt) (3%)
  - Comments (80%)
- Other unstructured data - Most appropriate incident image or photo (0.7%)

From this initial look at the number of fields, there is an immediate appreciation for why personnel are reluctant to record more data. While some of these fields (Project, Framework, Sector, Division, Incident Number, Postcode & Address) auto-populate, there are 23 other fields which require manual filling out. Some of the non-compulsory data fields, such as investigation summary and suggested improvements, have extremely low return rates. Meanwhile, the categorical data fields have so many different possible categories that their use becomes subjective and unclear (see results in Section 5.5.1). Because of the high proportion of incomplete field entries and risk of subjectivity elsewhere in the forms, the most pertinent information for learning from these events is contained within the text fields documenting the event itself.

The data used in this analysis consist of 14,882 injury, near miss and observation reports captured in 491 project locations, across 16 sectors in the UK. The number of data entries for each sector is shown in Table 5.2.

Table 5.2: Demographic of all available data

Sector	Number of Injuries	Total Number of Data
Airports	8	88
Commercial	1	69
Education	49	172
Environmental Services	12	42
Highways	1299	4155
Legacy (Infrastructure)	2	3
Legacy (Natural Resources)	32	91
Nuclear	223	405
Offices	9	20
Oil & Gas	211	1302
Power	235	435
Rail	1424	5559
Retail	18	153
Tunnels	1	4
Waste	71	268
Water	554	2116

NB. ‘Legacy’ sectors refer to inactive projects where the company has been required to return and perform remedial work to the asset.

5.3 Theory: Unstructured to structured data

Structuring the unstructured failure text data as a set of fundamental attributes is instinctively understandable and, as previously mentioned, this step can be used to increase the trust and explainability of further digital analytic or AI methods.

However, natural language processing (NLP) and machine learning (ML) are hugely diverse areas of research and therefore identifying an appropriate task type for the automatic extraction of attributes from the unstructured text is an essential first step. To inform this decision, existing work using NLP + ML pipelines to analyse text construction data is explored in sub-section 5.3.2.

However, to be able to sufficiently understand this exploration, select an appropriate set of methods for the required task and correctly interpret the results, there are several key ML concepts which must be understood. Therefore, these points of theory are first defined in sub-section 5.3.1.

5.3.1 Key ML concepts

This research is not a piece of data science or machine learning research. However, it does explore the use of data science/ML concepts for accessing the useful information within text failure data, and therefore it is essential to have an appreciation for some key ML concepts. ML is a vast area of research, therefore the concepts discussed here focus on classification type tasks - predicting a class - and do not explore regression tasks - predicting a number - as they are not relevant for this research.

Some of the definitions here were deemed relevant due to findings/decisions about the data to be analysed and required process selected. A ‘snapshot’ overview of these points is given here to aid understanding in the inclusion of each ML term definition. In the findings presented in Chapter 4, a need to better analyse and learning from text-based data from failures was highlighted; however, this qualitative piece also found a necessity to embed explainability and transparency into any AI analysis process. Therefore, I decided to adopt protocols developed and observed at the University of Colorado, Boulder (for example Tixier et al. (2016b)) to represent the text as a series of key attributes. From the existing literature in conjunction with the key considerations found from the initial qualitative analysis, I decided that identification of the attributes should be modeled as a text classification task.

Identifying the ML task

Identifying the ML task required is a key first step in consideration of relevant ML methods. Figure 5.1 is an adaptation of SciKit’s flowchart to select a relevant machine learning algorithm from their repertoire (Pedregosa et al., 2011). This can also be used as a decision guide to assess the type of task. In this flowchart, task-types are in bold: regression, classification, clustering and dimensionality reduction. Some key requirements for predicting a category are clear: (a) more than 50 data samples are required, and (b) the existence or creation of ‘labelled data’ needs to be considered.

A key appreciation for these methods is that they require a **structured** input, i.e. a numerical vector. For text input, an unstructured form, this text must be represented by a numerical vector. Deciding on this representation is an important part of the NLP process, described in sub-section 5.3.2.

Labelled data and supervised methods

For classification tasks, algorithms rely on supervised or semi-supervised methods. These both require at least some level of labelled data - that is examples of input data with the desired output attributes identified manually. Labels are the classes which are assigned to a specific data

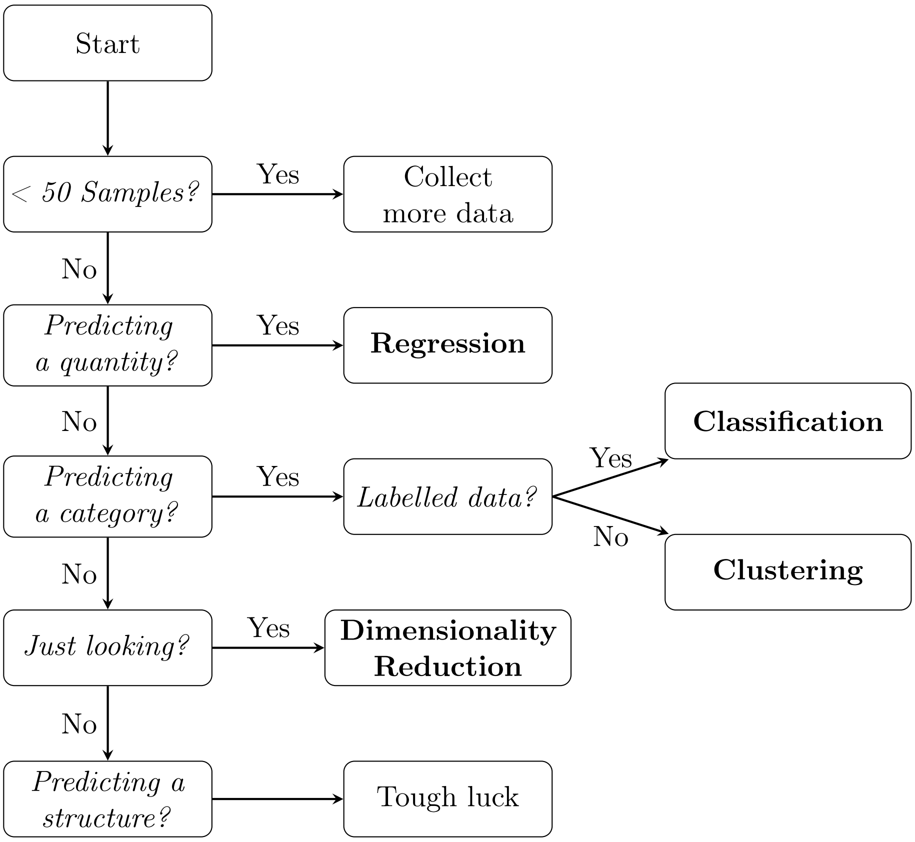


Figure 5.1: Adaption of 'machine learning algorithm selection diagram for scikit-learn' (Pedregosa et al., 2011)

point. From these labelled ‘training data’, ML algorithms can then learn the relationship between these inputs and the desired outputs. For semi-supervised methods, a lesser degree of labelling is required and some form of data clustering is used to label further.

Types of classification: binary, multi-class and multi-label

Classification tasks - where ML algorithms predict the class - can take the form of:

1. Binary: there are only two classes and the data belongs to one or the other
2. Multi-class: there are multiple classes and the data belongs to one of them
3. Multi-label: there are multiple classes and the data belongs to one or more of them

These tasks are progressively more complex moving from binary to multi-label. Additionally, the more classes which multi-class/multi-label are attempting to predict, the more complex the task.

In this research, more than one attribute could be present for each input (text safety event description), therefore, multi-class is inappropriate. The task could be considered as a multi-label task (where there is a class for each attribute), or a series of binary tasks (where a separate task is used for each attribute where the classes are “attribute present” and “attribute not present”).

Class imbalance

Class imbalance occurs when one class is more frequent than the other(s). In extreme cases, one class can dominate the dataset, especially visible in binary classification, where there are only two possible classes.

Class imbalance can confuse machine learning algorithms (Haixiang et al., 2017). A method employed for dealing with this is data sampling for the training dataset where the imbalance is addressed by artificially changing the ratio of the classes. Examples of methods employed include deliberately oversampling the positive counts or under sampling the negative ones.

In this research, imbalance is a key feature in identifying attributes in the text, as far more events occur without a specific attribute present than with it. Additionally, some attributes occur far more frequently than others (see sub-section 5.5.1 for attribute results).

Parameter vs hyperparameter

Machine Learning (ML) algorithms have internal parameters and external inputs. Internal parameter values are learnt and adjusted by the machine learning process; while external inputs, which are known as hyper-parameters, have to be input by the analyst. ‘Tuning’ or optimising hyper-parameters is an important step in optimising the model for a task.

To illustrate, fitting a polynomial curve to a series of 2D points can be considered a ML task. Polynomial equations, as in Equation 5.1, consist of a series of x terms each with a higher polynomial power than the last. In this example, the ML algorithm aims to optimise values of $A, B, C, D..Z$ to minimise the loss function between the line and the recorded points. These values, $A, B, C, D..Z$, are the parameters of the equation. However, the analyst inputs the value for n, the hyper-parameter in this case. A value of n which is too high will cause the line to over-fit, while a value which is too low will cause the line to under-fit.

$$y = A + Bx + Cx^2 + Dx^3 + ... + Zx^n \tag{5.1}$$

Model Performance Assessment

Assessing model performance relies on labelled data as it compares the expected output with the output predicted by the ML algorithm. The most widely used metrics for assessing model performance are recall, precision and F1.

- **Precision** measures whether a class predicted by the ML algorithm is correct or not.
- **Recall** measures the proportion of possible correct classes which were predicted.
- **F1** is a harmonic average of the previous two.

These metrics are calculated from values obtained in a confusion matrix which documents the expected class of the test data vs the predicted class. There is a trade-off between recall and precision - especially in imbalanced classes.

An example confusion matrix for 100 test cases is seen in Table 5.3. In this case, the class is actually present in 5 cases and not present in 95. However, the ML model has correctly predicted 4 cases where the class is present and incorrectly predicted an additional 5 cases.

Table 5.3: Confusion matrix example for classification

	Attribute (Predicted)	Not Attribute (Predicted)
Attribute (Actual)	4 (TP)	1 (FN)
Not Attribute (Actual)	5 (FP)	90 (TN)

where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative

These values are calculated as follows:

$$Precision = \frac{TP}{TP + FP} = \frac{4}{9} = 0.44 \tag{5.2}$$

$$Recall = \frac{TP}{TP + FN} = \frac{4}{5} = 0.8 \tag{5.3}$$

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision} = 2 \times \frac{0.8 \times 0.44}{0.8 + 0.44} = 0.57 \tag{5.4}$$

Another metric occasionally used is overall accuracy. This assigns equal value to predicting anything ‘correct’. For imbalanced classes, this value is dominated by the number of true negatives (TN). To illustrate, the example in Table 5.3 has $accuracy = (90 + 4)/100 = 94\%$. This gives a far inflated assessment of the results and could be interpreted to provide a misleadingly confident impression of the model performance. To illustrate this all the more, a ‘dumb’ model which simply always predicts the attribute as not present would achieve $accuracy = 95/100 = 95\%$, as it would correctly predict all those examples where the attribute is not present. The ‘dumb’ model would be shown to outperform the model from Table 5.3! Therefore, overall accuracy is unsuitable for model assessment in this research.

5.3.2 Previous use of AI for attribute extraction for safety reports

Previous research has used NLP (natural language processing) + ML (machine learning) pipelines for analysis of unstructured text construction data. As well as attribute-based analysis for safety incident descriptions, other analysis tasks have included retrieval systems and text classification. Table 5.4 contains a simplified overview of these examples, describing the task, the NLP representation used and the ML algorithm(s) investigated. These examples are listed in order of published year. This allows a view of the developing sophistication of methods over the years. All of those listed use safety incident reports as their input data - with the exceptions of Caldas and Soibelman (2003), Williams and Gong (2014) and Sun et al. (2020). These first two demonstrate early examples of NLP + ML in construction, while the last demonstrates a novel visualisation method.

Text Representation

Text representation refers to the manner in which the raw text data is converted into a structured vector which represents the text but can be used in downstream machine learning or analysis tasks. Only work by Tixier and colleagues (Tixier et al., 2016b; Tixier, Hallowell, and Rajagopalan, 2017) uses attributes to represent the text before the downstream task, i.e. retrieval, prediction or classification. In this case, they use a rule-based NLP method to extract an attribute set to represent the text.

Other research examples employ more generalised NLP methods (either Bag-of-Word or word embedding) to transform the raw text into a numerical vector representation which is then used directly in a downstream task. Bag-of-Words representations remain popular despite the increased semantic information encoded in word embedded vectors (discussed in sub-section 2.5.3). Despite the increased semantic information and complexity of the word embedded representation, research has not yet shown this translates into significantly improved task metrics. For example, in a piece of collaborative work, Baker, Hallowell, and Tixier (2020b) compared classification performance performance of both BoW and word embedding, and we found only a small improvement for complex deep learning methods.

However, in considering the requirement for a transparent and understandable method, use of an unintuitive numerical vector to represent the text is unsuitable. While for some applications and contexts, this ‘black-box’ approach may be sufficient; for learning from these data and increasing trust in these systems - especially for safety critical systems - this is a significant limitation. In light of this, representing the text as a set of attributes is adopted for this research. There are four ways previous research has attempted to extract attributes, or key concepts, from construction text: rule-based NLP extraction, keyword expansion, predictive region identification and NER-like identification. Each of these are now explored briefly, however, please note only the first had been published prior to this research commencing in 2016.

Table 5.4: NLP + ML construction literature

Reference	Task	Representation	Analysis algorithm(s)
Caldas and Soibelman, 2003	Classification of management documents	BoW+TF-IDF	NB, k-NN, Rocchio, SVM
Yu and Hsu, 2013	Retrieval of accident reports	BoW+TF-IDF	Vector similarity
Williams and Gong, 2014	Prediction of cost overruns	BoW+TF-IDF**	Riddor, K-Star, RBF neural nets
Tixier et al., 2016b + Tixier et al., 2016a	Prediction of safety outcomes	81 attributes	RF and Boosting
Tixier, Vazirgiannis, and Hallowell, 2016	Classification/retrieval of accident reports	Word vectors	Word Mover's Distance, k-NN
Tixier et al., 2016b + Chokor et al., 2016	Classification of accident reports	BoW+TF-IDF	Unsupervised clustering
Tixier, Hallowell, and Rajagopalan, 2017	Attribute clustering	81 attributes	Community detection, hierarchical clustering
Goh and Ubeynarayana, 2017	Classification of accident reports	BoW+TF-IDF	NB, k-NN, RF, LR, SVM
Zou, Kiviniemi, and Jones, 2017	Retrieval of accident reports	BoW*	Vector similarity
Kim and Chi, 2019; Moon et al., 2018	Retrieval of accident reports	BoW+TF-IDF*	Rule-based, CRF
Marzouk and Enaba, 2019	Classification of contractual documents	BoW+TF-IDF	Clustering
Zhang et al., 2019	Classification of accident reports	BoW+TF-IDF	NB, k-NN, RF, LR, SVM
Cheng, Kusoemo, and Gosno, 2020			
Zhong et al., 2020	Classification of accident reports	Word embedding	CNN, SVM, kNN, NB
Sun et al., 2020	Keyword extraction	BoW+TF-IDF with PoS	Clustering
Baker, Hallowell, and Tixier, 2020b	Classification of accident reports & Predictive region identification	BoW+TF-IDF and Word embedding	SVM, HAN, CNN

Representation Algorithms: Bag-of-Words (BoW), Term frequency-inverse document frequency (TF-IDF), Part-of-Speech (PoS)
Algorithm Acronyms: Naïve Bayes (NB), k-Nearest Neighbor (k-NN), Random Forest (RF), Linear Regression (LR), Support Vector Machine (SVM), Conditional Random Fields (CRF), Convolutional Neural Networks (CNN), Hierarchical Attention Networks (HAN).
*with word2vec and thesaurus implementation, **with bigrams.

Methods for attribute identification

Rule-based NLP extraction

The earliest example found of automated attribute case analysis, developed at the University of Colorado, USA, formed a set of features and a rule-based NLP method to extract these from text data. As presented in Tixier et al. (2016b), their features were categorised 'precursors' and 'outcomes'. These were developed by manual identification of attributes in 1280 injury reports and required 6 iterations to achieve their 95% agreement rate. The resultant attributes were seen previously in Table 5.1.

The 'precursors' had been further classed into 'upstream', 'transitional' and 'downstream' by Desvignes (2014). 'Upstream', 'transitional' and 'downstream' refer to the time period in which these attributes could be reasonably identified as a requirement/present at the worksite. 'Upstream' refers to during the design phase, 'transitional' refers to the pre-construction planning and 'downstream' refers to during construction i.e. at point of work risk assessment stage. All attributes are supposed to be identifiable BEFORE an incident occurs. However, the division of attributes into these three categories appears to be extremely subjective. For example, while it may be clear that 'working at height' could be identified for most situations at the design phase, the use of 'guardrails/handrails' may be a pre-construction decision rather than a design one. Equally, it could be argued that requiring 'nails' or 'screws' is identifiable at design phase.

The main benefit of this rule-based AI method is the extremely high performance rate achieved for attribute identification. Tixier et al. (2016a) reported 96% F1 performance (see sub-section 5.3.1 for full metric definition), which is unlikely to be achievable with general ML methods. However, this method is extremely resource intensive. Not only does it require labelled data (a requirement for all supervised machine learning tasks) to achieve this high-performance, they had to employ an extremely rigorous attribute labelling scheme for creation of the labelled data set - requiring 95% agreement between independent researchers - and had to place labels on individual words/phrases rather than across the entire text, known as annotation rather than simple labelling. Their rule-based approach also required iterative manual derivation of identification rules, another time-consuming task.

Additionally, this method can only identify the 80 'known' precursor attributes, i.e. their set of previously identified attributes. Ideally, it would be desirable to not only identify the known, or most common, attributes, but also identify those which are uncommon or novel. This limitation is a common theme among the methods explored here.

The final significant draw-back to this method - or any rule-based method - is that from the moment the process is complete, it becomes dated. With no way for the model to easily adapt to new ways of language use (either phraseology or grammar) or new attributes, the model is specific to the context in which it was created - geographically, temporally, and industry.

Keyword expansion

Another set of methods which could be employed to identify attribute presence is keyword searching. Zou, Kiviniemi, and Jones (2017) demonstrated the use of keyword expansion in a document retrieval task by using keyword expansion to find similarity between the search phrase and incident title. Keyword expansion is the process of defining a list of synonyms for a particular word in order to search texts for this entire list - not just the original keyword. In Zou, Kiviniemi, and Jones (2017), the keyword expansion to find synonyms was performed using a risk-related lexicon and WordNet (a large open-source lexical database).

Exemplars of this type of search can, of course, be found in internet search engines. While early search systems for large databases required exact keyword matching and employed complicated structured query logic (e.g. using Boolean logic to expand queries), this 'expert knowledge needed' approach was inappropriate for the general public wishing to search the web. To this end, Google now uses extremely sophisticated keyword expansion, finding synonyms which depend on the context of the query to help guide its search algorithm (Google, 2000). However, it is important to realise that the high performance of Google is dependent on its PageRank algorithm, which uses the links within the text it crawls through to perform a rank analysis for webpages, using the links as 'citations' to other pages which help determine the importance of the pages. This rank is then used in conjunction with the keyword search to rank the relevance of webpages to a user search input. Equally, Google's high speed is dependent on its 'pre-indexed' pages, containing defining information about each page it crawls - it does not search each page on the internet each time! These other factors mean that searching documents in this way, rather than webpages, is often far less effective.

While both these examples use keyword expansion to assess the similarity between the search phrase and the existing document database for document retrieval purposes, this keyword expansion method could be applied to find the existence of words/concepts in the text. In this way, attributes could be found within text data.

However, this method has several limitations:

1. It requires an existing set of attributes. Like the rule-based method, this method required a pre-defined set of attributes, or base words, to search for. This limits the model to looking for the 'known' attributes.
2. It does not have the ability to take ANY semantic context into account, for example, if using the attribute list from Table 5.1, the phrase 'no stairs were in place' would still return the attribute 'stairs' if performing a keyword search. With other methods, this may still be the case (false positive), however, with some of the other ML models, there is a chance that the model will encode the negative information in the sentence.
3. There is no way to link meaningful phrases to attributes, for example, the attribute 'work at height' may be signposted in the text by phrases such as 'at the top of [..]' or 'on the roof'. These phrases are completely unrelated, by synonyms, to 'work at height', but are still illustrative of the attribute.

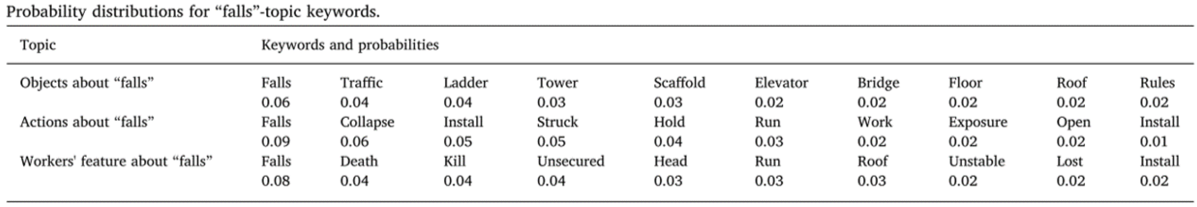
The advantages of this method over the previous rule-based method is that it requires no labelled data creation, or creation of complex identification rules. Although, it would require some phase of manual analysis to identify attributes (keywords). It can also be applied to new attributes by adding another word to the attribute list and re-running the algorithm - without requiring any adjustment to the core method.

Predictive region identification

Both Zhong et al. (2020) and my own collaborative paper (Baker, Hallowell, and Tixier, 2020b) use OSHA incident descriptions as raw text data input and then classify them using deep-learning methods. This research was published at the end of the research process. While Zhong et al. (2020) classify these documents into their incident type, e.g. 'fall from height', 'collapse of object', Baker, Hallowell, and Tixier (2020b) also classify them by incident severity, body part affected and injury type. In both these analyses, they do not use an intermediate attributes feature representation, i.e. they go directly from raw text as an input using NLP to create a numerical vector which is used for incident type classification. As mentioned, this introduces a level of obscurity to the method and limits the usefulness to industry professionals. To address this, both papers then go on to attempt to identify attributes by using different methods to find the most predictive parts of the text. The key assumption here is that worksite attributes are predictive of the safety event outcomes - an assumption which is backed by previous research, i.e. Tixier et al. (2016a).

In Zhong et al. (2020), they employ topic mining using a Latent Dirichlet Allocation (LDA) method to extract 'keywords' most predictive of each category. LDA uses the topic frequency in the total document set and word frequency in each topic to calculate the importance of each word to the topic. They suggest that these keywords illustrate connections in a similar manner to analysis of incident attributes. The keywords in their results are separated by objects, actions and workers features - although it is unclear how this distinction has been made: part-of-speech labelling or manual separation. An example of their results for a single category - falls - is shown in Figure 5.2.

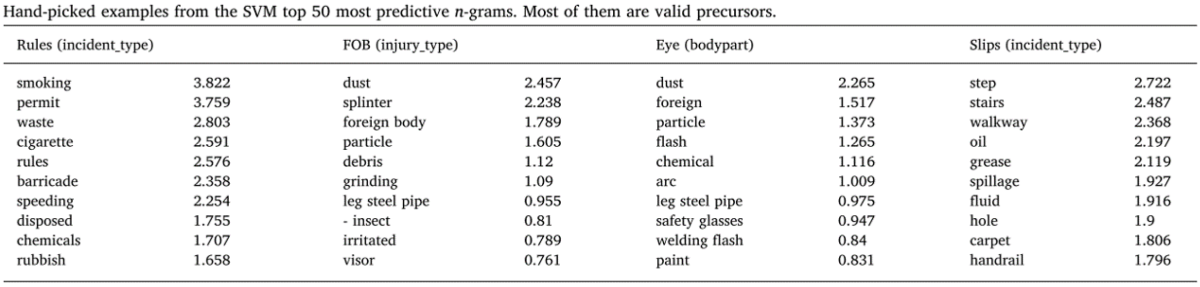
Figure 5.2: Example results for predictive keyword identification for 'falls' category (Zhong et al., 2020)



Probability distributions for "falls"-topic keywords.

Topic	Keywords and probabilities									
Objects about "falls"	Falls	Traffic	Ladder	Tower	Scaffold	Elevator	Bridge	Floor	Roof	Rules
	0.06	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.02	0.02
Actions about "falls"	Falls	Collapse	Install	Struck	Hold	Run	Work	Exposure	Open	Install
	0.09	0.06	0.05	0.05	0.04	0.03	0.02	0.02	0.02	0.01
Workers' feature about "falls"	Falls	Death	Kill	Unsecured	Head	Run	Roof	Unstable	Lost	Install
	0.08	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.02

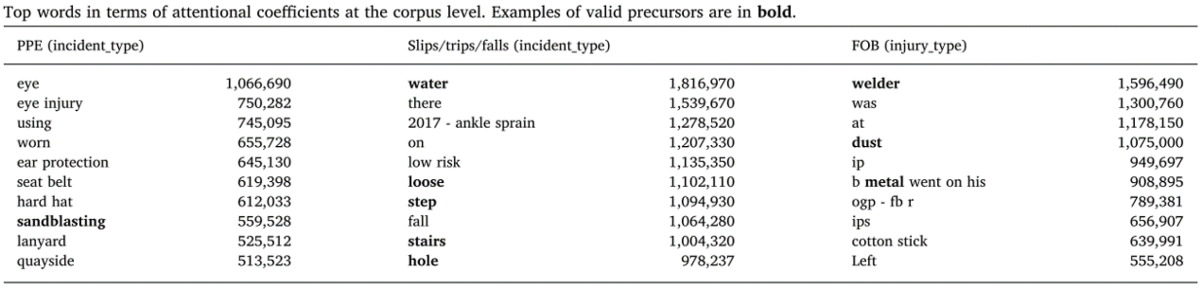
Meanwhile, Baker, Hallowell, and Tixier (2020b) identified predictive words or phrases by two different methods. They applied word saliency (analysis quantifying the significance of each word to the outcome classification) to identify the most predictive words. Also, hierarchical attention networks (HANs), one of the deep learning methods experimented with, also has built in attention coefficients for words and phrase which can also be used to extract key predictive words or phrases. Examples of these results for HAN and SVM (two of the machine learning methods used) are shown in Figures 5.3 and 5.4.



Hand-picked examples from the SVM top 50 most predictive *n*-grams. Most of them are valid precursors.

Rules (incident_type)		FOB (injury_type)		Eye (bodypart)		Slips (incident_type)	
smoking	3.822	dust	2.457	dust	2.265	step	2.722
permit	3.759	splinter	2.238	foreign	1.517	stairs	2.487
waste	2.803	foreign body	1.789	particle	1.373	walkway	2.368
cigarette	2.591	particle	1.605	flash	1.265	oil	2.197
rules	2.576	debris	1.12	chemical	1.116	grease	2.119
barricade	2.358	grinding	1.09	arc	1.009	spillage	1.927
speeding	2.254	leg steel pipe	0.955	leg steel pipe	0.975	fluid	1.916
disposed	1.755	- insect	0.81	safety glasses	0.947	hole	1.9
chemicals	1.707	irritated	0.789	welding flash	0.84	carpet	1.806
rubbish	1.658	visor	0.761	paint	0.831	handrail	1.796

Figure 5.3: Example results for SVM predictive n-grams (Baker, Hallowell, and Tixier, 2020b)



Top words in terms of attentional coefficients at the corpus level. Examples of valid precursors are in **bold**.

PPE (incident_type)		Slips/trips/falls (incident_type)		FOB (injury_type)	
eye	1,066,690	**water**	1,816,970	**welder**	1,596,490
eye injury	750,282	there	1,539,670	was	1,300,760
using	745,095	2017 - ankle sprain	1,278,520	at	1,178,150
worn	655,728	on	1,207,330	**dust**	1,075,000
ear protection	645,130	low risk	1,135,350	ip	949,697
seat belt	619,398	**loose**	1,102,110	b **metal** went on his	908,895
hard hat	612,033	**step**	1,094,930	ogp - fb r	789,381
sandblasting	559,528	fall	1,064,280	ips	656,907
lanyard	525,512	**stairs**	1,004,320	cotton stick	639,991
quayside	513,523	**hole**	978,237	Left	555,208

Figure 5.4: Example results for HAN attentional regions (Baker, Hallowell, and Tixier, 2020b)

These methods have a significant advantage over the previous two discussed; they do not require a manual analysis to identify the attribute set. While they require a labelled data set for the outcome categories, i.e. incident type, the identification of predictive regions is not dependent on a pre-set list of attributes and does not require a labelled set of data for attributes. This represents a large decrease in manual input/time and also provides the ability to identify the 'unknown unknown', i.e. it can identify predictive attributes which would not be picked up by manual labelling - perhaps due to personal biases.

However, the largest drawback with this method is that many of the predictive regions or words refer to sections of the incident description which describe the outcomes or are not valid precursor attributes (worksite attributes identifiable before the incident occurs). This is evident in all the example result tables given - even the handpicked examples in Table **??** which aims to show examples of categories where the identified regions are mostly precursors. For 'FOB' (foreign object injury), the method has identified 'irritated' and 'foreign body' as predictive phrases which clearly refer to resultant injury suffered due the foreign object, rather than pre-identifiable attributes for the worksite. Zhong et al. (2020)'s result in Figure 5.2 are even more illustrative of this phenomenon as 'falls' is the most predictive word for the incident category of 'falls', followed by 'collapse'.

Another limitation is that by using the predictive region to identify key attributes for the incident means that this method is specific to this data type - safety incidents - and cannot then be applied to other text data which does not have these outcomes (although other failure data could be used to discover the most predictive attributes for that specific outcome, e.g. quality NCR data could be used to find the most predictive attributes for rework vs leave-as-is).

Identification akin to Named-entity recognition (NER)

During the course of of this research project, the UK Health and Safety Executive (HSE) commenced a programme of research, entitled the 'Discovering Safety Programme' which began to investigate analysis of text data from their RIDDOR (Reporting of Injuries, Diseases and Dangerous Occurrences Regulations) reports from construction projects for a similar purpose: identifying key factors and attributes for the purpose of learning from these failures. Once again, this data set only contains events which resulted in an injury.

Their research team at the text-mining institute at the University of Manchester framed this as a named-entity recognition (NER) problem. NER is normally used to identify proper nouns, company names, or places within text. However, the research team proposed that similar algorithms could be employed to identify sets of attribute types. They define seven attribute types: protection measure, body part injured, harmful consequence, construction activity, equipment, physical environment, and hazard. These sets were created to be consistent with existing safety risk onotologies in UK construction (BSI, 2018; Zhang, Boukamp, and Teizer, 2015).

In the few years the 'Discovering Safety' programme has been on-going, the research team at the University of Manchester has produced a set of 600 annotated reports and recently published a paper on their annotation guide on labelling the data (Thompson et al., 2020). Having attended several meeting with their research group, both at HSE Buxton and with the text-mining group at the University of Manchester, it is clear that this annotation method was an extremely intensive process. While similar to the annotation conducted by Tixier et al. (2016b), a key difference is that Thompson et al. (2020) annotated attribute type rather than individual attributes. Therefore, a further stage of clustering will be required later to group equivalent individual attributes. This annotation was performed by a group of text-mining researchers. For NER, the individual spans of text indicating the presence of a particular attribute type must be annotated. This is a complicated process and also was found to be quite subjective in some cases, with an exact inter-annotator agreement (IAA) rate of only 66%. IAA can be calculated for exact span matches (where annotators identify the same attribute type and the exact same span) or relaxed span match (where annotators identify the same attribute type and overlapping spans). In this case, using the relax span metric increased IAA to 79%.

To briefly investigate this method, I co-supervised a dissertation project for a student undertaking a taught MSc in Speech and Language Processing, based at the School of Philosophy, Psychology and Language Sciences. My co-supervisor, Prof Bonnie Webber, is a computational linguist based at the University of Edinburgh's School of Informatics. As Thompson et al. (2020) published their annotated safety incident descriptions along with their paper, Murray (2020) used this pre-annotated dataset and applied various implementation of NER algorithms to identify attribute spans. For attribute spans, that is spans which identify attributes identifiable at the engineering planning stage, three attributes types from (Thompson et al., 2020) were used: construction activity, equipment, and physical environment. It is key to realise that, used in this way, the methods explored found the span type, but did not group these into usable individual groups for further analysis. Four methods were investigated for the NER task:

First, a baseline was established using a keyword expansion-type method which, for the reasons already explained, performed poorly, with F1 = 0.31.

Second, a multiclass SVM was trained to assess probabilities that each word embedding belonged to each class, selecting the most probable class for that word. This method achieved F1 = 0.54. Murray (2020) found that using construction specific word embeddings gave a modest increase in performance.

Third, Conditional Random Field (CRF) tagger was used to identify the spans. For this, each input included the following features: word, part-of-speech, shape, parent dependency, head word. Also, the word was input into the previous SVM model and the class membership probabilities as features for each token. This model achieved an average F1 = 0.67 across the three entity types.

Finally, BERT-large (Bidirectional Encoder Representations from Transformers) language model (Devlin et al., 2018) was fine tuned on the same data and combined with a simple softmax classifier to identify the attribute group. BERT is a pre-trained word-embedding type model, designed at Google to be context specific and achieves extremely high performance on a variety of natural language tasks. Additionally, as the model is trained on sub-word tokens, it can cope with misspelled words in the original data. In his research, Murray (2020) gives the example of 'slabas' being correctly identified as a member of the physical environment attribute group, as a misspelling of 'slabs'. In this case, however, it struggled to identify legal BIO sequences, which resulted in a performance slightly lower than the CRF model, with F1 = 0.66. In the future, this could be easily adjusted with an additionally check on span legality before assessing the performance.

This NER method of attribute identification has some notable advantages. It does not have a set list of attributes to identify - allowing the 'unknowns' to be extracted from the attribute group sets. It can be trained on relatively little labelled data - 600 sentences - and achieve a performance on-par with the IAA.

However, by finding only attribute groups, a further method would need to be applied to find equivalent attributes, for example, to group 'cherry picker' and 'manlift' for further analysis. Additionally, to produce the annotated data set required a large investment in time for the University of Manchester team, and proved extremely difficult to achieved a good agreement rate between researchers (79% relaxed span agreement).

Classification methods

As seen in Table 5.4, there are several previous examples of text classification - where the entire piece of text is assigned a class - using construction text. While a labelled data set is still required, expert linguistic knowledge is not required as the labels can apply to the entire text, rather than the lexical span. Additionally, outside the realm of construction failure, there are numerous examples of text classification which can be used to inform the method development. For these reasons, I decided to pursue this line of enquiry for identifying attributes within the safety event descriptions.

Interestingly, just because the other examples of NLP + ML pipelines for construction did not use attributes as their classes for the classification task, does not mean that their findings are irrelevant for this research. Lessons can be taken from these classification tasks to develop a method for identifying attributes via text classification with similar data type (i.e. text descriptions of safety incidents). Additionally, in comparing the results of the classification tasks for this extremely similar data and taking into account the differences in complexity for the class types, these examples can provide some comparison examples to include in the discussion later in the research.

As a brief overview, the classification tasks listed here have large differences in terms of prediction complexity and methodology. For example, in Baker, Hallowell, and Tixier (2020b) (and Tixier et al. (2016a)), multiple injury characteristics are predicted using ensemble machine learning methods, including 7 (4) types of injury, 5 (6) body parts and 6 (2) severity levels. In Baker, Hallowell, and Tixier (2020b), they also predict 6 incident types and experimented with model stacking. Meanwhile, Esmaeili, Hallowell, and Rajagopalan (2015) classify only a binary severity outcome (fatal vs. non-fatal) with a simple logistic regression model and Goh and Ubeynarayana (2017) predict only three categories (no accident, minor accident or major accident) using five standalone machine learning algorithms.

An initial experiment using classification to identify key concepts in short technical text was undertaken as a collaboration with The Welding Institute (TWI). For this investigation, they had a set of document abstracts with keywords assigned and wanted to be able to automate labelling new abstracts. This is a similar task to predicting attributes from safety event descriptions: using a short technical description to extract key concepts from a pre-defined list. The advantage of first testing this method on the TWI data was that the data was already labelled. Labelling data, as discussed, represents a large investment in human resource and therefore exploring the potential of this method prior to creation of a data set labelled for text classification (which is not transferable to the annotated version required for NER-like analysis) was a rational decision. The final report for this investigation is included as Appendix E with key findings included next.

Initial NLP + ML test - TWI Investigation

As mentioned, an initial experiment using classification to identify key concepts in short technical text was undertaken as a collaboration with The Welding Institute (TWI). The aim of this piece of work was to explore the different classification ML algorithms and natural language processing (NLP) to automate keyword selection for TWI abstracts. This is a similar task to predicting attributes from safety event descriptions: using a short technical description to extract key concepts from a pre-defined list.

Initial exploration of the TWI abstract data distribution found that of the 1945 unique keywords used to label the abstracts, 913 – nearly half all keywords - occurred in less than 0.1% of the abstracts. Only 183 keywords occurred in more than 2% of abstracts. This high proportion of low occurring keywords will prove to also be a feature of worksite attributes.

Some key findings were:

1. Deep learning required greater than 1000 positive examples for benefits to occur.
2. Imbalanced class distribution in this data proved to be a hindrance to achieving high accuracies, especially recall values.
3. Oversampling the minority class to minimum 10% positive examples significantly increased the F1 performance. However, for Decision Tree methods (i.e. Decision Tree and Gradient Boost) this oversampling method decreases the precision of the model, while significantly increasing the recall.
4. Gradient Boost ensemble is the highest performing algorithm for the resampled data and is recommended for use in implementation. It slightly outperforms SVM, SVM bagging and the Decision Tree models. These algorithms significantly outperform Naïve Bayes and kNN. (The specifics of these algorithms are introduced in Section 5.4.2.)
5. SVM and SVM bagging algorithms are more precise than Gradient Boost at higher positive proportion keyword assigned.

While many defining features of this text classification task are similar to extracting attribute, there are a few differences. The most significant one, which must be acknowledged, is that these abstracts were written by individuals who all possessed a high level of literacy, being the authors of technical reports or academic papers. This contrasts to the various levels of written literacy of personnel responsible for filling out incident reports on construction projects. In particular, for NLP, misspellings and mixed up homophones hinder the analysis. Therefore, in applying the findings of this initial investigation to the construction project data, methods which can adjust for spelling and grammar error should be considered.

To conclude, this investigation demonstrated the potential of modelling the task of keyword prediction as a series of binary text classification tasks. The similarity of this task to attribute prediction indicated that this is a suitable method. Several key findings are pertinent for development of the attribute prediction model; these will be referred to during the Method sub-section.

5.4 Method: Unstructured to structured data

There are several NLP + ML pipelines which could have been suitable for the text classification task chosen to predict attribute presence. A two-step process was created, following from the results from the TWI exploratory study using text classification from key words and adapting from protocols developed and observed at the University of Colorado, Boulder (for example, Tixier et al. (2016b)). Here, the specific steps used for the two-part method are: (1) developing a set of work-site attributes and labelled data via manual analysis, and (2) employing supervised text classification ML algorithms to predict the presence of these attributes.

5.4.1 Data labelling method: Unlabelled to labelled data

For supervised ML processes, a set of 'learning data' which is already labelled with the correct classes (in this case, construction attributes) is required. Two methods of labelling have been presented in other research: whole-document classes and span annotation.

For whole-document classes, the entire text is labelled as belonging to a class. Meanwhile, for span annotation, a 'class span' is labelled which refers only to the word or word span describing the class. Whole-document labelling is appropriate for the text classification task chosen to carry out the automatic extraction of attributes, while span annotation is required for Named-Entity Recognition (NER) type tasks, such as previously discussed in Section 5.3.2, where the classifier is classifying for individual works (or spans) rather than the whole text.

In considering span annotation, label classes could either consist of the individual attributes, i.e. 'hammer', as in Tixier et al. (2016a) or attribute groups, i.e. 'tool', as in Thompson et al. (2020). The main advantage of the latter is that the classes are fewer, therefore both the labelling exercise and machine learning task are arguably simpler. However, a further step would be required to group equivalent attributes before use in a downstream task. Additionally, this annotation scheme is unusable for whole-document classification of attributes, as the individual attributes are not labelled. In contrast, if the span classes are the individual attributes, these attributes could be extracted for whole-document classification, however the annotation task would be extremely complex.

I chose to undertake whole-document text labelling rather than span annotation, as this was suitable for extracting individual attributes via text classification and for using engineers as the labelling team rather than linguists/trained annotators. This decision was based upon the assumption that even well-trained teams would struggle to consistently identify attribute spans. This assumption (made near the beginning of the research journey) was recently backed up in independent research when Thompson et al., 2020 implemented span annotation of attribute groups and found achieving high inter-annotation agreement (IAA) extremely difficult. In their research, they achieved a relaxed IAA - where annotators identify the same attribute type and overlapping spans - of 79%.

A set of safety reports were manually labelled with their construction attributes by four researchers at the University of Edinburgh, within the School of Engineering, using Microsoft Forms to collate the data. They also labelled these data with safety outcome attributes, such as 'immediate cause' and injury details, to enable a quality check on the data.

It is vital to recognise the impact of these researchers on the development of the attribute dataset. By using personnel familiar with construction and engineering, rather than linguists, they should have been able to more accurately identify the pertinent construction information in the text. However, previous experience of construction could also have resulted in unconscious bias where individuals have preconceived notions about what is important on construction sites.

Table 5.5 demonstrates an output from the labelling process, linking unstructured text to their work attributes. Note, as mentioned, that the entire text was labelled as containing an attribute, rather than labelling an 'attribute span'.

Table 5.5: Examples of safety event descriptions and labelled attributes

IP slipped down temporary steps. The steps were wooden and wet which made them slippery.	Objects: Stairs Activity: Moving around Worksite environment: Slippery surface
IP was cutting the old safety barrier with a cut off saw. He went to step over the barrier to make a new cut when the saw, which was still running, slipped causing an abrasion to his left thigh.	Objects: Barrier, Powered Saw, Sharp Edge Activity: Cutting Worksite environment: None

As discussed in Section 5.3.2, Desvignes (2014) defined precursor attributes as attributes which are identifiable before an incident occurs and contribute to the incident occurring (see Table 5.5). Analysis by (Tixier et al., 2016b) using these precursor attributes also considered safety incident outcome attributes and safety incident category, both using NLP+ML to automatically extract this information from the text (Tixier et al., 2016b) and also prediction of these outcomes using the precursor attributes (Tixier et al., 2016a). Outcome attributes - body parts, severity and injury type - are generally also captured separate from the text data on the form as categorical data (drop-down selection). This is also true of 'safety incident category', such as slip-trip-fall. In the industry dataset used here, this was formed of two levels - a high level category (1Incident Type') and a secondary level category ('Incident Sub Category') which can be more accurately described as an 'immediate cause' category, where 'immediate cause' is the action or decision at the point of the safety event which caused the event to occur.

This research used Desvignes (2014)'s existing set of precursor attributes as a starting point, initially reconsidering the categories of precursor attributes. The three temporal divisions - upstream, transitional and downstream - along with existing categorisations were removed and the attributes were separated into four categories: objects - materials, tools and machinery; actions - actions being undertaken which contributed to the incident; and worksite descriptors - defining features of the workspace or area of incident; and personnel descriptors - defining features of the people involved. While undertaking the annotation and analysis, it became clear that the personnel descriptors consisted of 'immediate causes' rather than attributes identifiable before the incident occurred, therefore those attributes identified were moved. Outcome and safety category attributes were initially a combination of those from the multiple choice on the form and those from Tixier et al. (2016a).

MS Forms was used to capture and combine the annotation of the novel data. The form consisted of a set of 'pre-suggested' options, as well as the option to use a free text box to add a different attribute. It is acknowledged that using 'pre-suggested options' could have resulted in bias for these easier-to-select attributes (the full list of these is in Appendix). The following questions were included on the form:

1. Enter unique ID
2. Not relevant data entry

Site attributes.
These are descriptors or attributes which can be identified BEFORE the incident occurs.

3. Which of the following OBJECTS contribute to the incident occurring?
4. Which of the following WORK SITE DESCRIPTORS contribute to the incident occurring?

5. Which of the following PERSONNEL DESCRIPTORS contribute to the incident occurring?
6. Which of the following ACTIONS contribute to / occur while the incident occurring?

Consequence.
These are descriptors of the incident consequence.

7. Incident ‘immediate cause’
8. Injury Type
9. Body part

Three iterations of the labelling exercise were undertaken: (1) test software and attribute types, (2) initial trial with correlation check and (3) main labelling exercise. The ‘pre-suggested’ options were revised at each step.

For the first iteration, an initial set of the ‘precursor’ categories were used, as taken from Tixier et al. (2016a). However, in application of an attribute-based method for UK construction rather than US, annotators found that the terminology differed, e.g. wrench vs spanner, lumber vs timber, and that the list did not cover all situations in the novel dataset. However, in considering the UK data, the safety outcome categories were considered consistent although some terminology differs eg laceration vs open wound. So while these attributes were used as a starting point, they are not suitable for use in the UK and required refining.

The second iteration revised the ‘pre-suggested’ options following discussions in light of the annotators’ findings from a small independent labelling samples. Three revisions of the ’pre-suggested’ options took place. A sample of 56 report description were then labelled by all annotators to calculate the inter-annotation agreement using the method outlined in Nowak and Rüger (2010).

The final stage involved independent annotation of over 3000 safety event descriptions to produce a set of labelled data and a final list of attributes in the data. The results of this exercise are in sub-section 5.5.1.

It is important in labelling data that a representative data set is labelled. However, considering the limitations on manual labelling resource, priorities had to be made. To this end, the four most frequent sectors in Table 5.2 - Water, Rail, Highways and Oil&Gas - representing 88% of all the data were prioritised. A demographic of the project demographics labelled are in Table 5.6.

Table 5.6: Demographic of labelled data

Sector	Number of Injuries	Total Number of Data
Water	114	320
Rail	333	1584
Highways	408	1086
Oil&Gas	20	244
Other infrastructure	4	8

5.4.2 AI method for automatic attribute prediction

Having obtained a labelled set of data, a supervised NLP + ML pipeline can be implemented for text classification to predict those attributes present in the safety event descriptions. Figure 5.5 demonstrates the NLP + ML pipeline adopted for this research. The steps in this pipeline are:

1. **Construction Safety Text Data**: collection of the data as detailed in sub-section 5.2.
2. **Labelled data**: data labelling exercise, as detailed in previous sub-section (5.4.1), where entire text is labelled with attributes.
3. **Data division**: in line with good data science practice, a proportion of the labelled data is set aside for result testing and evaluation - here, 10% was set aside. These data are not used in any hyper-parameter optimisation or training tasks. Here, the data are divided before the NLP transformation in order to better replicate the operational reality of such a model - where 'new' data are not included in the creation of the natural language vector space dictionary. In keeping with this philosophy, the data are divided into 5 equal-sized sets (i.e. k-fold where k=5) and each set sent separately to the NLP transformation such that each validation data transformation is based on the token dictionary from the 80% training data.
4. **Natural Language Processing (NLP)**: the natural language data are converted into a numerical vector to be used on the ML algorithms. Here, a TF-IDF vector space (aka 'Bag-of-Words') representation is used. The specific steps for this are elaborated upon in this sub-section.
5. **NLP Transformation to TF-IDF Vector**: non-training data are transformed using the same steps as shown in (4), however, the vector space is based on the training data. This means that the dimensions (tokens) are already set, so new tokens/words are not included except as an 'other' dimension.
6. **Classification Machine Learning**: This classification took the form of a series of binary classification tasks, where a piece of text either belonged to the class containing an attribute or to the class not containing the attribute. This is due to the number of attributes and the complexity if modelled as a multi-label task. The specific steps for this and the algorithms investigated are detailed at the end of this sub-section.
7. **Predict attributes**: the best-performing trained algorithms can then predict attributes for new data. N.b. there is a binary classification algorithm trained per attribute.
8. **Downstream tasks - knowledge discovery**: some examples of these are explored in Chapter 6.

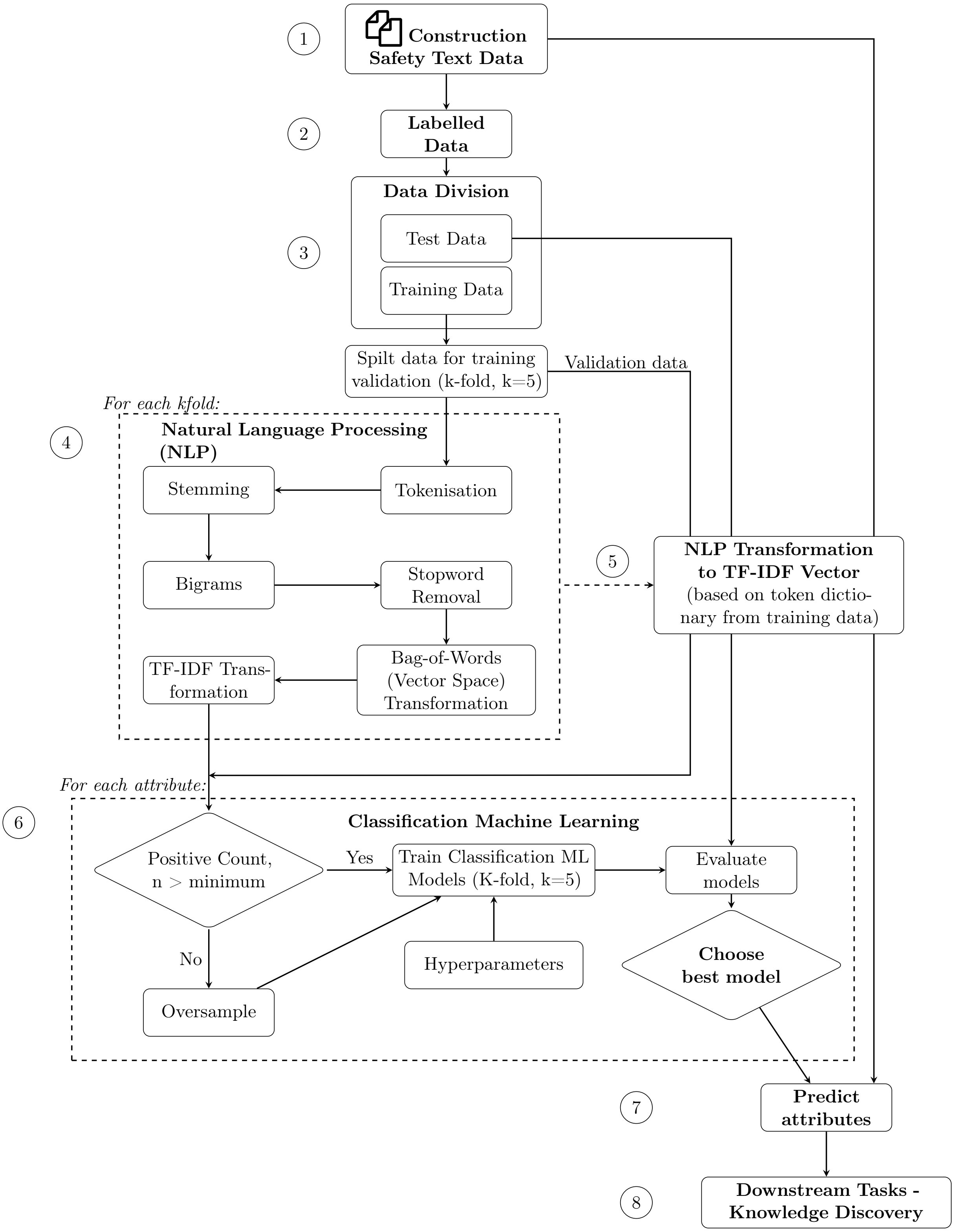

1
Construction Safety Text Data
2
Labelled Data
3
Data Division
Test Data
Training Data
Spilt data for training validation (k-fold, k=5)
Validation data
For each kfold:
4
Natural Language Processing (NLP)
Stemming
Tokenisation
Bigrams
Stopword Removal
TF-IDF Transformation
Bag-of-Words (Vector Space) Transformation
5
NLP Transformation to TF-IDF Vector (based on token dictionary from training data)
For each attribute:
6
Classification Machine Learning
Positive Count, n > minimum
Yes
No
Train Classification ML Models (K-fold, k=5)
Evaluate models
Oversample
Hyperparameters
Choose best model
7
Predict attributes
8
Downstream Tasks - Knowledge Discovery

Figure 5.5: Workflow

NLP Transformation Steps

1. Tokenisation: Tokens were created by splitting text on whitespace and punctuation.
2. Stemming: To decrease vocabulary length and integrate some semantic relationships into the model, the lexical stem of each token was extracted using the Snowball algorithm (Porter, 2001). For example, 'management' and 'managing' would both map to 'manag'.
3. Bigrams: To mitigate against word order loss, bigrams (pairs of words) which occur more than 5 times in the training data were found and included as tokens. For example, 'circular saw'. Less frequent and larger phrases were not included as this would increase the vector length and sparsity to such an extent that it becomes difficult to fit any model.
4. Stopword removal: At this stage, stopwords (words which are deemed not to add semantic meaning to text), punctuation and numbers were removed.
5. Bag-of-Words (Vector Space) transformation: Vector space transformation involves creating a dictionary of the tokens from the training data, where each token is included as a separate dimension. Pieces of text are transformed into this vector space by counting the numbers of times each token is present in the text. This creates a long, sparse vector.
6. TF-IDF transformation: TF-IDF transformation scales each token frequency (i.e. number of token counts) in the original vector by log(inverse document frequency) i.e. log(number of documents in total / number of documents containing the token). These logarithmically scaled word counts identify defining tokens for the document (Jones, 2004). This transformation can be considered standard for investigations using NLP vector space representations.

NB. These parameters, e.g. vocabulary and TF-IDF transformation coefficients, are calculated on the training data then applied to the test and validation set.

ML Algorithms

For prediction of attributes from the safety event description text, each attribute was considered independently. Binary classification algorithms were trained using the TF-IDF text representation with their associated classification for that attribute.

Four base classification algorithms were investigated. Deep learning algorithms were not used due to the low data numbers as the initial TWI investigation found that >1000 positive examples were required for benefits to occur. This was later backed by other research, such as Baker, Hallowell, and Tixier (2020a), where deep learning methods had minimal improvement over TF-IDF + SVM.

The four algorithm types investigated were: Naïve Bayes, k-Nearest Neighbour, Decision Tree and SVM (Support Vector Machines). Two ensemble methods - gradient boosting for Decision Trees and bootstrap aggregating (Bagging) for SVM - were also applied to the final two algorithms respectively. Tables 5.7 to 5.12 introduce these algorithms. The descriptions within them are based heavily on those included in Baker, Hallowell, and Tixier (2020a).

The relevant hyper-parameters associated with these algorithms are also coarsely optimised. As stated by Bottou, Curtis, and Nocedal (2018), "optimization is one of the foundations of machine learning"; this includes not just optimisation during the training process but optimisation of the hyper-parameters. Optimising algorithm hyper-parameters has a significant effect on the performance results. In this investigation, coarse optimisation was applied as the aim was to uncover the effects these hyper-parameters caused the models for this task, rather than to find the absolute optimisation.

Table 5.7: Naïve Bayes

<table>
<tr><th>Algorithm</th><th>Naïve Bayes</th></tr>
<tr><td><b>Short Description</b></td><td>This method uses the observed token (word) frequencies in the training data and uses the (Naïve) assumption that all of these frequencies are independent. This allows the application of Bayes Theorem to calculate both the probability that the attribute is present and the probability that the attribute is not present. The most likely situation is the predicted as true.</td></tr>
<tr><td><b>Illustrative example</b></td><td><b>Bayes Theorem</b>
$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$
Where $P(A) =$ the probability of the attribute,
$P(B) =$ the probability that the text is composed,
$P(B|A) =$ the probability of the text given that the attribute is present,
$P(A|B) =$ the probability of the attribute given the text – this is what we want to calculate in new cases.
NB. For a new sentence, P(B) is calculated by the observed token frequencies (e.g. *P('this is a sentence') =* $P(\text{'}this') \times P(\text{'}is') \times P(\text{'}a') \times P(\text{'}sentence')$)
<b>Worked Example:</b>
<table>
<tr><th>Number of training examples
n = 100</th><th>Attribute 'Stairs'</th><th>Not Attribute 'Stairs'</th></tr>
<tr><td></td><td>30</td><td>70</td></tr>
<tr><td>Step</td><td>5</td><td>15</td></tr>
<tr><td>Rust</td><td>2</td><td>20</td></tr>
<tr><td>Staircase</td><td>10</td><td>5</td></tr>
</table>
In this example, the observed frequencies are recorded above. These frequencies are used to calculate the probabilities. Given a simple example, where the text contains only the tokens 'step', 'rust' and 'staircase':
$P(A) = \frac{30}{100} = 0.3$ and $P(\overline{A}) = \frac{70}{100} = 0.7$
$P(B) = P(step) \times P(rust) \times P(staircase) = \frac{20}{100} \times \frac{22}{100} \times \frac{15}{100} = \frac{66}{10000} = 0.0066$
$P(B|A) = \frac{5}{30} \times \frac{2}{30} \times \frac{10}{30} = \frac{1}{270} = 0.0037$
$P(B|\overline{A}) = \frac{15}{70} \times \frac{15}{70} \times \frac{20}{70} \times \frac{5}{70} = \frac{15}{3430} = 0.00437$
Therefore, $P(A|B) = \frac{P(A) \times P(B|A)}{P(B)} = 0.168$ while $(P(\overline{A}|B) = 0.463$
As $P(A|B) < P(\overline{A}|B)$ then the model would predict that the text is not assigned the attribute 'stairs'.</td></tr>
<tr><td><b>Why this algorithm was selected</b></td><td>The TF-IDF vector space representation of the text description numerically emphasises defining tokens for each piece of text. If it is assumed that these defining tokens (word) are associated with the defining attributes for the work situation, then by employing Naïve Bayes, a statistical model which uses these adjusted frequencies as input, the attribute presence should be predicted.</td></tr>
<tr><td><b>Why this algorithm may be unsuitable</b></td><td>The assumption of token independence is incorrect - words depend on those around them. In this context, attributes may be related to words which may only refer to the attribute in combination with other words. Also, there exists a large amount of noise in these TF-IDF representation - i.e. words which are defining to the piece of text but do not relate to the attributes.</td></tr>
<tr><td><b>Python Implementation</b></td><td>sklearn.Naïve_bayes.MultinomialNB</td></tr>
</table>

Table 5.8: k-Nearest Neighbour

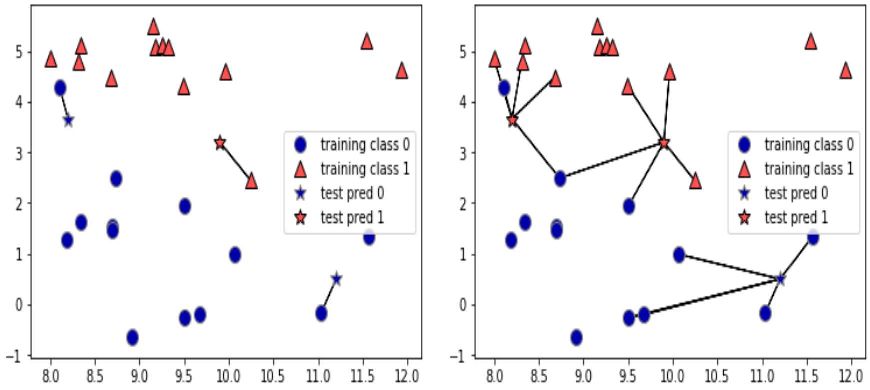
Algorithm	k-Nearest Neighbour
Short Description	This classification method uses vector distance to identify the closest example in the training set, then adopts this example's classification. If k>1, an average is used.
Illustrative example	These two figures illustrate kNN prediction for k = 1 and k = 4 for a binary 2-D example.  Image credit to Le (2018).
Why this algorithm was selected	The TF-IDF vector space representation of the text description numerically emphasises defining tokens for each piece of text. If it is assumed that these defining tokens (words) are associated with the defining attributes for the work situation, text representations relating to a certain attribute should then fall 'closer' in the multidimensional space to each other than those representations not relating to that attribute as the defining token dimensions dominate the position.
Why this algorithm may be unsuitable	Once again, there exists a large amount of noise in these TF-IDF representation - i.e. words which are defining to the piece of text but do not relate to the attributes. Additionally, the high number of dimensions and low dimensional values hinder the model.
Python Implementation	sklearn.neighbors.KNeighborsClassifier

Table 5.9: Decision Tree

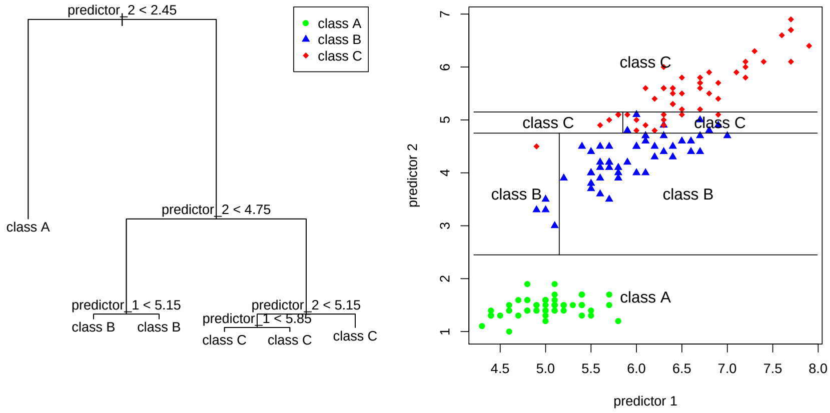
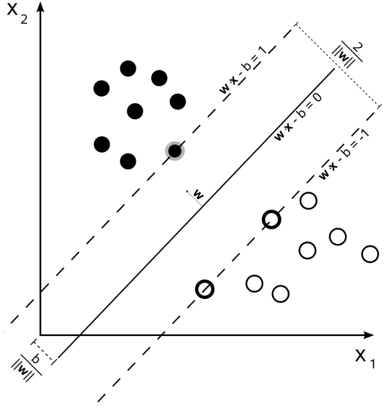
Algorithm	Decision Tree
Short Description	Decision trees aka CART (classification and regression tree) algorithms split the data using binary ‘queries’, repeating this until each end point only contains one data class. The ‘split’ or branches are created by maximising the ‘split goodness’ criterion, for example, Gini index.
Illustrative example	This example shows for a 3-class classification with 2-dimensions how a CART algorithm divides the data to predict a class.  predictor_2 < 2.45 class A predictor_2 < 4.75 predictor_1 < 5.15 class B class B predictor_2 < 5.15 predictor_1 < 5.85 class C class C class C ● class A ▲ class B ◆ class C class C class C class C class B class B class A predictor 2 1 2 3 4 5 6 7 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 predictor 1 Image from Baker, Hallowell, and Tixier (2020a)
Why this algorithm was selected	This method is often used in high accountability industries, such as medicine and finance, as it is easily interpretable and can be interrogated to explain why a certain decision was made. In this case, findings from Chapter 4 demonstrated the need for explainability. Additionally, as stated in Baker, Hallowell, and Tixier (2020a), “CART decision trees are able to capture complex nonlinear high-order interactions among predictors, scale well with the number of predictors and observations, and are relatively robust to outliers and irrelevant predictors. However, they often need to be grown very large to accurately represent the training data.”
Why this algorithm may be unsuitable	As stated in Baker, Hallowell, and Tixier (2020a), “there are two main negative side effects to growing a large decision tree: (1) poor generalization to unseen observations (overfitting) and (2) high variance, as the lower parts of the trees are very sensitive to changes in the training data. From the perspective of the bias-variance framework, where $error = bias + variance$, deep decision trees are ***low bias-high variance*** models.”
Python Implementation	sklearn.tree.DecisionTreeClassifier

Table 5.10: SVM

Algorithm	Support Vector Machines (SVM)
Short Description	SVM relies on graphical divisions to separate classes of data. In 2-D, this could be represented as a line best dividing the classes.
Illustrative example	This 2-D example shows a hyperplane separating the two classes. The best separating hyperplane is the line that separates the two groups of points with the greatest possible margin on each side. Training the SVM, that is, finding the best hyperplane, comes down to optimizing the $\vec{w}$ and b parameters.  X2 X1 w·x - b=1 w·x - b=0 w·x - b=-1 2/‖w‖ b/‖w‖ w As stated in Baker, Hallowell, and Tixier (2020a), "because in practice, points may not all be separable (e.g., due to outliers), when searching for the best separating hyperplane, the SVM is allowed to misclassify certain points. The tolerance level is controlled by a parameter traditionally referred to as C in the literature. The smaller C, the more tolerant the model is towards misclassification. C plays a crucial *regularization* role, i.e., it has a strong impact on the generalization ability of the SVM. Indeed, for large values of C (low misclassification tolerance), a smaller-margin hyperplane will be favored over a larger-margin hyperplane if the former classifies more points correctly, at the risk of overfitting the training data. On the other hand, small values of C will favor larger-margin separating hyperplanes, even if they misclassify more points. Such solutions tend to generalize better. Optimising C can, therefore, have a large effect on the performance of this model."
Why this algorithm was selected	SVM has been proven to produce good results in a variety of different tasks, including text classification.
Why this algorithm may be unsuitable	A general drawback to this method is how the time taken to train scales with the training data volume, known as time complexity. Finding the support vectors scales quadratically with the number of training examples n. In practice, time complexity is the main limitation of SVMs compared to ensemble CART methods.
Python Implementation	sklearn.linear_model.SGDClassifier which is equivalent to linear SVM when the default loss function ('hinge') is selected. Stochastic Gradient Descent (SGD) refers to the optimisation technique used to train the algorithm. As stated in the scikit, "SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing."

Bagging - definition from Baker, Hallowell, and Tixier (2020a)

The **B**ootstrap **agg**regat**ing** (bagging) method (Breiman, 1996a) was introduced as a way to reduce variance and therefore overfitting. The bagging procedure consists in training many base models in parallel on bootstrap samples of the data. A bootstrap sample is obtained by randomly selecting observations with replacement from the original training set until a dataset of the same size is obtained.

Approximately one third of the observations are not expected to be present in each bootstrap sample, as the probability of not selecting a given observation with replacement from a sample of size n is $(1 - \frac{1}{n})^n$, which tends to $\exp(-1) \approx \frac{1}{3}$ when n tends to infinity. These observations compose what is called the 'out-of-bag' (OOB) sample (Breiman, 1996b). Since the bootstrap sample is of same size as the original dataset, it follows that for a large number of observations, each bootstrap sample is expected to contain about two thirds of unique examples, the rest being duplicates.

If applied to Decision Tree (CART) models, this causes each tree in the ensemble to become an expert on some specific domains of the training set. In this way, it is possible to take advantage of the low bias of deep decision trees while reducing their high variance. Bagging thus creates an ***ensemble of local experts***.

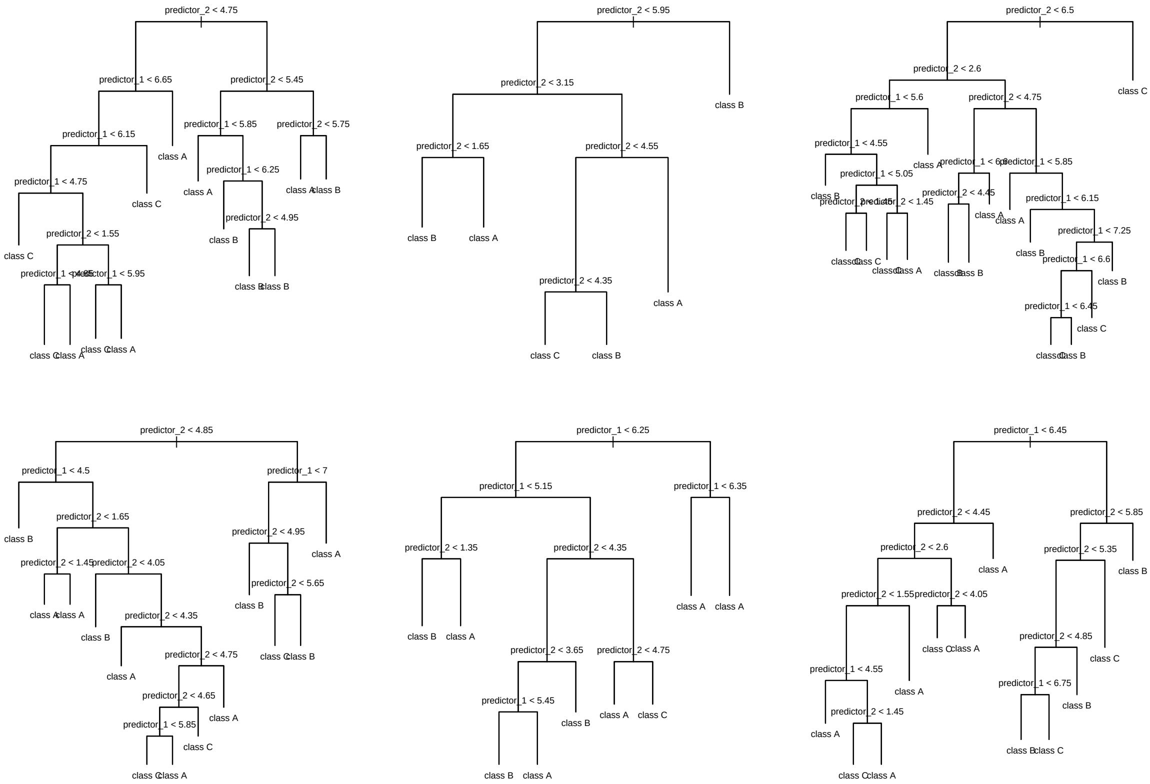


Figure 5.6: Example of a bagged ensemble of decision trees. Each tree in the ensemble is grown on a bootstrap sample of the original data. The large differences in the tree structures highlight well the high-variance nature of decision trees.

Thus, at prediction time, there will be a significant amount of beneficial disagreement among trees (see Fig. 5.6). By aggregating the predictions of all trees in the ensemble via majority voting, one obtains a model with significantly less variance than a single tree. Such a model generalizes much better, while still having almost the same low bias. This approach is known as ***perturb and combine*** - a slightly evolved version of this ensemble is known as **random forest**.

Despite being a significant improvement over CART, bagged ensembles are less interpretable. Also, by definition of CART, only those variables yielding the greatest decrease in node impurity are selected at each split. Consequently, all the trees in the bagged ensemble have quite similar upper structures, and tend to generate correlated forecasts, which reduces the disagreement among trees and prevents the maximal reduction in variance from being achieved. In this case, where the number of variables is large, this behaviour would limit the usefulness of random forest. Therefore, gradient boosting is applied in this research.

For SVM, bagging has the same effect of creating local experts which then aggregate to create a strong predictor. This has the effect of reducing overfitting and has been proven to greatly outperform the base SVM model, such as in Kim et al. (2002).

Table 5.11: Gradient Boosting: High Variance-Low Bias Ensemble of Decision Trees

Algorithm	Gradient Boosting
Short Description	This ensemble method regularises the base algorithm by training many shallow examples and averaging the results. Regularising algorithms reduces their potential for overfit. In this case, the boosting algorithm (Freund and Schapire, 1997) adds weak high bias-low variance base models (shallow decision trees which predict only slight better than guessing, i.e. the final leaves still contain quite mixed classifications) in sequence, repeatedly reducing the bias of the entire sequence.
Why this algorithm was selected	This model has been proven to achieve significant improvements over the base CART model by addressing some of the major drawbacks, i.e. overfitting and reducing the bias of the model to the training data.
Why this algorithm may be unsuitable	The ensemble version of the CART model is less explainable than the base model.
Python Implementation	sklearn.ensemble.GradientBoostingClassifier

Table 5.12: SVM Bagging

Algorithm	SVM Bagging
Short Description	Also known as bootstrap aggregating, bagging is an ensemble method where the base model is trained several times on different bootstrap samples of the training data, and the results averaged.
Why this algorithm was selected	To reduce possible overfitting in the base SVM model, especially in light of significant noise in the vector space representation of the input text.
Why this algorithm may be unsuitable	Bagging increases the runtime of the model and, as previously covered for the SVM method, time scales poorly with data volume. This makes this method unsuitable for large data sets.
Python Implementation	sklearn.ensemble.BaggingClassifier and sklearn.linear_model.SGDClassifier

Python algorithm implementations used and hyper-parameters explored

For this research, data analysis was implemented using Python programming language. Module versions used were:

- scikit-learn = 0.21.2
- pandas = 0.24.2
- numpy = 1.16.4
- nltk = 3.4.4
- matplotlib = 3.1.0
- gensim = 3.4.0

The next set of tables (Table 5.13 to Table 5.17) provide details of the hyper-parameters explored for each ML algorithm. Recall from Section 5.3 that a hyper-parameter is an input value which the analyst specifies for the algorithm, as opposed to the parameters which are trained by the algorithm itself. Most algorithms have more than one hyper-parameter input. For this research, a "coarse grid" approach is used. This is where there are large intervals between the hyper-parameter values being input ("coarse") and every possible combination of these hyper-parameters is attempted ("grid"). This approach is used to reveal the task's sensitivities and patterns to the hyper-parameter selection. If being optimised for an operational environment, a much finer grid would be appropriate.

Table 5.13: Naïve Bayes Hyper-parameters

Naïve Bayes	
Hyper-parameter grid	alpha: 0.1 (default), 1, 10
Alpha	An issue with the Naïve Bayes algorithm occurs when a word appears in the test text which wasn't observed in the training data. If $P(word) = 0$, then $P(text) = P(B) = 0$, $P(A\|B) = 0$ as well as $P(\overline{A}\|B) = 0$. Smoothing techniques can be employed to overcome this. The most common, known as 'additive smoothing', simply adds an observed count (or counts or part of a count) to every word in the final vocabulary. 'Alpha' for Naïve Bayes implemented from scikit in Python is an additive smoothing constant and controls the degree of smoothing employed.

Table 5.14: k-Nearest Neighbour Hyper-parameters

k-Nearest Neighbour	
Hyper-parameter grid	n_neighbours: 1,5 (default), 10
n_neighbours	The number of neighbours hyperparameter determines the number of nearest points which should be considered for averaging to determine the class. If n=1, only the closest point is considered, making the model vulnerable to outliers. While is n is large, examples which are not representative as they are far away may be included.

Table 5.15: Decision Tree Hyper-parameters

Decision Tree	
Hyper-parameter grid	max_depth: 1,2,3,4,5,10
max_depth	This hyperparameter sets a maximum depth for the tree. Low values will create weak estimators, where the final leaves are not single classes; meanwhile high values risk overfitting the data.

Table 5.16: Gradient Boosting Hyper-parameters

Gradient Boosting	
Hyper-parameter grid	max_depth: 1, 2, 3 (default), 5, 10 max_features: None (default), log2, sqrt learning_rate: 0.1 (default), 0.5, 1, 10, 100 n_estimators: 10, 50, 100 (default), 150
max_depth	As for Decision Tree, this hyper-parameter sets a maximum depth for the tree.
max_features	The number of features to consider when looking for the best split. If ‘None’, max_features = n_features. If ‘sqrt’, then max_features=sqrt(n_features). If ‘log2’, then max_features=log2(n_features).
learning_rate	“Increasing learning rate shrinks the contribution of each tree by learning_rate. There is a trade-off between learning_rate and n_estimators.”
n_estimators	“The number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number usually results in better performance.”

Table 5.17: SVM Hyper-parameters

SVM - SGD Implementation	
Hyper-parameter grid	Alpha: 0.0001 (default), 0.0002, 0.0003, 0.0004 Penalty: ‘l1’, ‘l2’ (default) loss: ‘hinge’ (default), ‘log’
Alpha	‘Alpha’ for SGD implementation of SVM is indicative of the regularisation constant, C. For this model,
Penalty	Penalty function ‘l2 is the standard regularizer for linear SVM models. l1 might bring sparsity to the model (feature selection) not achievable with l2.’
Loss	Changing the loss in SGD is not strictly optimisation of the hyperparameters for SVM, rather it changes the base model. When loss is ‘hinge’, an SVM model is used. When loss is ‘log’, a logarithmic descent model is used. This therefore investigates whether SVM is outperformed by logarithmic descent.

5.5 Results: Unstructured to structured data

5.5.1 Data labelling results

Developing work attributes

Incident work attributes were considered to be observable features of the construction site and activity prior to any incident occurring. These followed the definitions set out in PAS 1192-6:2018 “Specification for collaborative sharing and use of structured Health and Safety information using BIM” by identifying actions/activity, objects (materials, tools and machinery) and site environment descriptors. Originally, personnel descriptions were also included, such as “incorrect PPE” or “fatigued”; however, these attributes and those which emerged in this category during the labelling exercise were more accurately described as “immediate causes” rather than precursor attributes which could be observed before the incident occurred. Therefore, these were excluded as precursor attribute.

In total, 3491 incident reports from 28 infrastructure projects in 10 sectors were labelled - of which, 3244 reports were unique. Those which were duplicated, from the second iteration of the labelling exercise, are used to calculate inter-annotator agreement, adopting the method presented in Nowak and Rüger (2010). In this method, confusion matrices are constructed for the agreement of each annotator, as in Table 5.18. These tables show each annotator’s labelled data compared to another’s labelled data. For example, (a) A1-A2 provides information for when annotator 1 (A1) and annotator 2 (A2) agreed or disagreed while labelling. In 175 instances, they both labelled a text description as containing a specific attribute (1). In 6915 instances, they both both labelled a text description as not containing a specific attribute (0). However, in 190 instances, one annotator labelled a text description as containing a specific attribute while the other did not.

Table 5.18: Annotator Agreement Confusion Matrices for All Attribute Categories

	1	0
1	175	97
0	93	6915

(a) A1-A2

	1	0
1	161	111
0	88	6920

(b) A1-A3

	1	0
1	172	100
0	90	6918

(c) A1-A4

	1	0
1	156	112
0	93	6919

(d) A2-A3

	1	0
1	213	55
0	49	6963

(e) A2-A4

	1	0
1	158	91
0	104	6927

(f) A3-A4

From these, the ‘total agreement’ and ‘agreement’ can be calculated. ‘Total agreement’ measures how many of the labels and non-labels the annotators agreed on. In imbalanced data, this can be dominated by the agreement that a label is not present - as is the case here. Meanwhile, ‘agreement’ only considers those where at least one of the annotators has labelled the attributes as present. These values are shown in Table 5.19 and the average across all annotators given in the caption. Meanwhile, Table 5.20 provides an average agreement over the annotators dis-aggregated into attribute type group.

$$TotalAgreement = \frac{TP + TN}{TP + FN + FP + TN} \tag{5.5}$$

$$Agreement = \frac{TP}{TP + FN + FP} \tag{5.6}$$

where TP = True Positive - both annotators labelled attribute present (1), TN = True Negative - both annotators labelled attribute not present (0), FP = False Positive - annotators disagreed, and FN = False Negative - annotators disagreed

Table 5.19: Annotator Agreement

	A2	A3	A4
A1	0.974	0.973	0.974
A2		0.972	0.986
A3			0.973

(a) Total Agreement, average = 0.975

	A2	A3	A4
A1	0.479	0.447	0.475
A2		0.432	0.672
A3			0.448

(b) Agreement, average = 0.492

Table 5.20: Average Agreement by Attribute Type

Attribute Group	No. Attributes in Group	Average agreement	Average total agreement
Actions	13	0.452	0.958
Objects	43	0.451	0.976
Site descriptions	15	0.399	0.976
Person descriptions	5	0.409	0.914
Immediate cause	33	0.639	0.985
Body Part	14	0.660	0.990
Injury	7	0.464	0.964
Average		0.492	0.975

The findings demonstrate a moderate agreement (0.4-0.6). It is widely accepted that the more categories there are, the harder it is to get annotation agreement. Therefore, this agreement level is sufficient for the extremely high number of categories annotated here.

Higher levels of annotator agreement could be achieved through many successive iterations, as in Tixier et al. (2016b). However, I chose not to pursue this here. This was due to two key reasons. First, increased bias - especially if there is a power imbalance in the discussion groups. Second, it is not representative of what would be achieved in industry.

In development of the attributes set, 553 work attributes were identified in the labelling exercise. Many of these attributes were similar and, therefore, the next iteration combined attributes which had the same, or extremely similar, semantic meaning. Examples include 'animal', 'rat' and 'mouse' attributes, identified during the labelling exercise, all map to a single 'animal' attribute for analysis.

In total, 250 unique work attributes were identified. This is a much higher number than the 30 previously identified by Tixier et al. (2016b). Additionally, only 60 attributes (listed in Table 5.21) occurred in 1% or more of the safety descriptions.

This high proportion of infrequent attributes is indicative of the complexity of a construction site environment, which often sees specialist tools, materials and activities. Although most construction personnel could probably name the frequent activities and their main components, terminology differs across the country, increasing the complexity of the labelling task. These factors can also affect the performance of text classification, as discussed in the next sub-section.

Table 5.21: List of precursor attributes occurring >1% of labelling safety event descriptions

Attribute Type	Attributes			Number
Actions	Cutting Cleaning Lifting/pulling/ manipulating (manual)	Driving Exiting/entering Striking/stripping (i.e. formwork/ shuttering)	Using (a tool) Lifting (by machinery) Walking/moving around	9
Objects	Airborne particles Alarm Cabin Cable Concrete Crane Door Electrical Source Formwork Gate Guardrail Hand size pieces Hazardous substances	Heavy object Heavy vehicle High fence Light Vehicle Lumber/timber Machinery Manlift Mobile phone Mud Object at height Object on the floor Piping Pressure systems	Rebar Scaffold Sharp edge Small machinery Small particles Stairs Steel sections Storage tank Unpowered hand tool Vegetation Vehicle (unspecified) Water source	38
Environment	Adverse weather (storm, rain) Exclusion zone Poor Housekeeping Uneven surface Working at height	Congested/confined work space Insufficient edge/fall protection Railway Unstable support / surface	Excavation Slippery surface Vehicle Movement Zone Wind	13

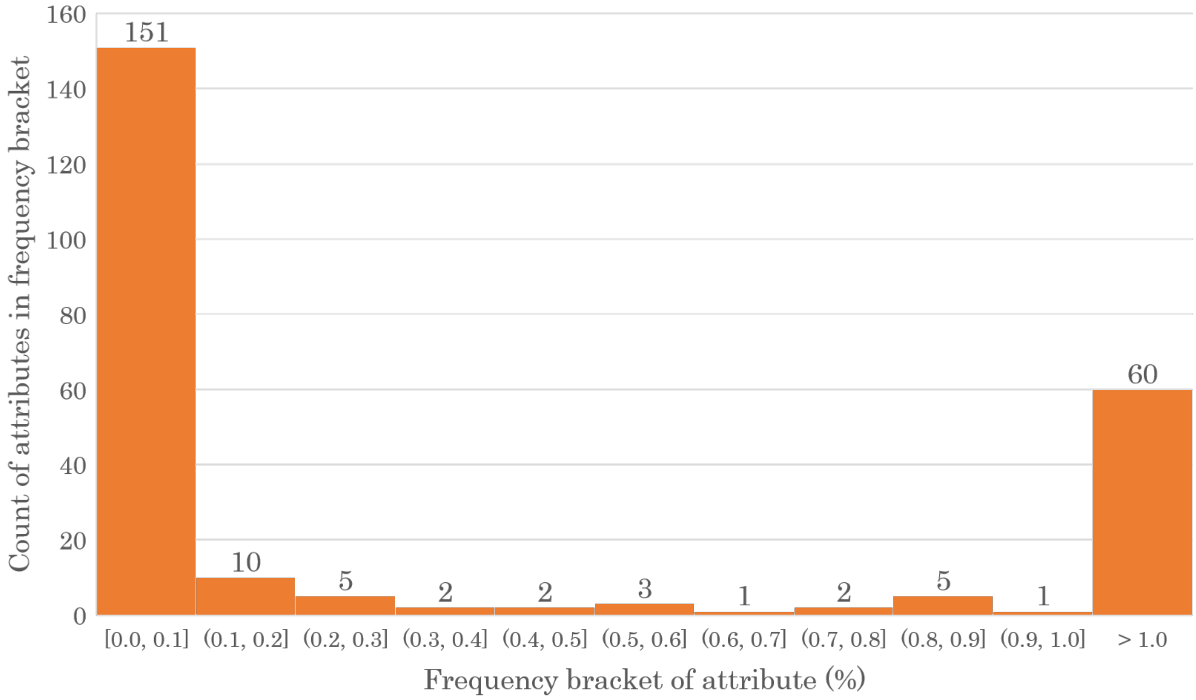


Figure 5.7: Attribute frequency in percentage of labelled data

On the other hand, the 60 attributes which occurred in over 1% of the safety event descriptions fully described 81% of the descriptions, i.e. 81% of annotated descriptions were not annotated with any of the less frequent attributes. Additionally, 113/250 of the unique attributes identified only occurred once in the annotation set. These factors prompt discussion of how granular attributes need to be to be representative of the text data and useful for analysis. This discussion is picked up in Chapter 7, Section 7.1.

Table 5.22: Attribute groups frequency distribution

	Precursor			Outcome		
	Action	Object	Site	Cause	Body part	Injury
0 - 0.1	27	107	14	18	13	30
0.1 - 0.2	1	11	0	4	1	3
0.2 - 0.3	1	3	0	1	1	2
0.3 - 0.4	0	2	2	1	1	0
0.4 - 0.5	0	1	1	3	1	5
0.5 - 0.6	0	2	1	1	1	1
0.6 - 0.7	0	0	0	4	1	0
0.7 - 0.8	0	2	0	1	0	0
0.8 - 0.9	0	5	1	1	2	0
0.9 - 1	0	1	0	3	0	0
>1	9	38	13	32	10	8

Examining safety event immediate causes and outcomes

During the labelling process, the safety event immediate cause and outcomes were also examined. While these were listed as multiple choice categories in the original data, researchers labelling the data also examined the free-text descriptions to see if this information concurred with that given in the text descriptions of the safety event.

Of the total 3491 safety event reports labelled, annotators disagreed with the immediate cause given by the report writer in 580 (16.6%) of these. When labelling these data, annotators were given the instruction to only re-label this category if they truly believed that the person on site had got the wrong category - the site person was to be given benefit of the doubt in most cases. Therefore, this high proportion demonstrates the subjectivity of this multiple choice category. This is discussed further in discussion in relation to the appropriate collection of data for learning.

Additionally, in 90/2075 (4.3%) near-miss reports, it was found that an injury occurred but was not recorded as an accident report. Of these, 33 were unspecified injuries (i.e. “IP hurt his hand”) and 14 were reports of natural causes of illness (e.g. flu, cold, headache). The rest were evenly distributed among the other minor injury types. The mis-reporting of injuries could indicate a reluctance to flag an investigation or adversely affect the project statistics for ‘minor’ injuries, or they could represent a need for further clarity in the training about how to log these minor events.

5.5.2 Predicting attributes from free-text

In presenting the results of the attribute prediction, first an overview of the test performance scores are shown for four scenarios for each model investigated. Individual factors are then investigated in further detail, including the volume of available training data and the detailed effects of oversampling. Results for each of the three attribute types - objects, actions and environment - are dis-aggregated to explore whether the type of attribute affects the model. Finally, the coarse grids of hyper-parameters for each algorithm (introduced in Tables 5.13 to 5.17) are explored.

As mentioned, only attributes which were observed in 1% or more of the training data set were considered. This is partly due to the inability of the models used to deal with the extremely imbalanced data classes, and partly because there is a high chance that they are completely absent from the test data.

Tables 5.23 and 5.24 show the F1, Precision and Recall performance scores for the 5-fold validation runs and test (unseen) data respectively. Higher values are shown in green, while lower values are increasingly red. The highest values for each are underlined and in bold. For 5-fold validation, an average standard deviation is indicated (calculated using the average variance).

Hyper-parameters and oversampling are coarsely optimised, rather than finely optimised, as the aim of this exercise was to investigate the effects of these factors upon the model, not find the absolute 'best' result. For the test data, the models have been trained (and optimal hyper-parameter set identified) on the entire training set.

Each table shows four stages of implementation:

1. Default hyper-parameters with no oversampling of the data
2. Coarsely optimised hyper-parameters with no oversampling of the data
3. Default hyper-parameters with coarse oversampling optimisation
4. Coarsely optimised hyper-parameters with coarse oversampling optimisation

As can be seen, across all four scenarios for both 5-fold validation and the final test data, Naïve Bayes, kNN and Decision Tree were outperformed by Gradient Boosting and SVM algorithms.

Oversampling had a significant effect on increasing the recall scores of all the models. For SVM (and to lesser extent k-Nearest Neighbour), the positive effects of this were traded-off against the slight negative effect this had on precision. Tuning hyper-parameters had a positive effect for all performance metrics for all models in the 5-fold validation, however, for the final test, using the previously optimised parameters had a negative effect. This is discussed further at the end of this section, however, is (in short) due to the coarse nature of the optimisation used and the high variance of the model.

The standard deviation, σ, demonstrates the high variance nature of Knn, decision trees and SVM as the spread is high. However, when comparing the performance scores for 5-fold data and test using default parameters, we actually see an increase. This is likely due to the increase in training data available for the full test run. Remember that in k-fold, where $k = 5$, 80% of the data is used for training the models while 20% is used for validation scores. The effect of training data volume is explored next.

For the remainder of this section, unless explicitly stated, the results shown are for the final test set with the model having been trained on the full training data and any optimised hyper-parameters selected from the k-fold optimisation task.

Table 5.23: Overview of attribute prediction performance scores - 5-fold validation

Default hyper-parameters	**F1**		**Precision**		**Recall**	
	Average	σ	Average	σ	Average	σ
Naïve Bayes	0.115	0.062	0.287	0.202	0.080	0.047
K-Nearest Neighbor	0.243	0.104	0.478	0.253	0.182	0.086
Decision Tree	0.375	0.119	0.388	0.150	**0.390**	0.145
Gradient Boosting	0.381	0.123	0.453	0.152	0.366	0.148
SDG Support Vector Machine	**0.382**	0.133	0.579	0.223	0.311	0.132
SDG SVM Bagging	0.324	0.116	**0.587**	0.228	0.244	0.107

Optimised hyper-parameters	**F1**		**Precision**		**Recall**	
	Average	σ	Average	σ	Average	σ
Naïve Bayes	0.115	0.062	0.287	0.202	0.080	0.047
K-Nearest Neighbor	0.292	0.105	0.395	0.154	0.264	0.121
Decision Tree	0.408	0.127	0.450	0.167	0.409	0.154
Gradient Boosting	0.471	0.141	0.526	0.186	**0.480**	0.172
SDG Support Vector Machine	**0.473**	0.138	0.633	0.223	0.410	0.137
SDG SVM Bagging	0.441	0.138	**0.642**	0.225	0.371	0.144

Default hyper-parameters	**F1**		**Precision**		**Recall**	
w oversampling	Average	σ	Average	σ	Average	σ
Naïve Bayes	0.329	0.122	0.387	0.191	0.328	0.136
K-Nearest Neighbor	0.300	0.106	0.352	0.143	0.352	0.152
Decision Tree	0.447	0.119	0.430	0.152	0.505	0.156
Gradient Boosting	**0.497**	0.131	0.494	0.163	**0.552**	0.149
SDG Support Vector Machine	0.431	0.146	**0.580**	0.215	0.375	0.152
SDG SVM Bagging	0.438	0.143	0.561	0.202	0.391	0.149

Optimised hyper-parameters	**F1**		**Precision**		**Recall**	
w oversampling	Average	σ	Average	σ	Average	σ
Naïve Bayes	0.393	0.118	0.399	0.151	0.465	0.158
K-Nearest Neighbor	0.323	0.106	0.360	0.127	0.353	0.142
Decision Tree	0.471	0.130	0.470	0.170	0.521	0.161
Gradient Boosting	0.546	0.142	0.554	0.180	**0.590**	0.154
SDG Support Vector Machine	**0.550**	0.142	**0.584**	0.181	0.569	0.159
SDG SVM Bagging	0.546	0.141	0.579	0.178	0.567	0.168

Table 5.24: Overview of attribute prediction performance scores - final test results

Default hyper-parameters	**F1**	**Precision**	**Recall**
	Average	Average	Average
Naïve Bayes	0.171	0.311	0.132
K-Nearest Neighbor	0.268	0.454	0.212
Decision Tree	0.448	0.479	**0.445**
Gradient Boosting	**0.463**	0.588	0.414
SDG Support Vector Machine	0.430	**0.661**	0.349
SDG SVM Bagging	0.387	0.625	0.300

Optimised hyper-parameters	**F1**	**Precision**	**Recall**
	Average	Average	Average
Naïve Bayes	0.171	0.311	0.132
K-Nearest Neighbor	0.301	0.383	0.281
Decision Tree	0.442	0.503	0.422
Gradient Boosting	**0.444**	0.544	**0.431**
SDG Support Vector Machine	0.430	**0.661**	0.349
SDG SVM Bagging	0.434	0.613	0.356

Default hyper-parameters	**F1**	**Precision**	**Recall**
w oversampling	Average	Average	Average
Naïve Bayes	0.374	0.417	0.382
K-Nearest Neighbor	0.323	0.380	0.373
Decision Tree	0.458	0.456	0.485
Gradient Boosting	**0.521**	0.505	**0.589**
SDG Support Vector Machine	0.464	**0.608**	0.410
SDG SVM Bagging	0.470	0.560	0.434

Optimised hyper-parameters	**F1**	**Precision**	**Recall**
w oversampling	Average	Average	Average
Naïve Bayes	0.423	0.422	0.506
K-Nearest Neighbor	0.317	0.346	0.357
Decision Tree	0.446	0.461	0.465
Gradient Boosting	0.518	0.512	0.577
SDG Support Vector Machine	**0.542**	**0.541**	**0.584**
SDG SVM Bagging	0.516	0.509	0.560

Set against other recent research, such as explored in sub-section 5.3.2, these F1 performance values may seem comparatively low. For example, Zhong et al. (2020) achieved an average F1 = 0.59 using SVM classification on word embedded text representations, achieved via word2vec with skip-gram. Meanwhile, in my own collaboration piece, Baker, Hallowell, and Tixier (2020a) achieved F1 = 0.72 using TF-IDF representation and SVM on a set of 6 incident types. These authors also achieved marginally higher F1 averages than those reported here by using deep learning classifiers. However, these results are not directly comparable for the following four identified reasons.

Firstly, these predictions had significantly fewer categories. For example, Zhong et al. (2020) predicted only 11 categories. Also, these were not attributes but incident categories, e.g. 'electrocution', 'falls'. Fewer categories mean that outlier accuracies can more significantly affect the average. In this case, 'electrocution' predicted with F1=0.92 brings the average from 0.46 to 0.59.

Secondly, having fewer categories may indicate a lower class imbalance. This is also indicated in the category types. 'Incident category' or 'type' tends to be a multiple-choice option on safety incident report forms and is compulsory in most cases. This means that not only are there fewer categories, but every incident must contain at least one of them. Additionally, Baker, Hallowell, and Tixier (2020a) employed a tailored oversampling factor for each class. Oversampling was shown to have a significant effect on the performance results - as further explored here- and finely optimising this could have significant advantages.

Furthermore, the data used in both previous papers mentioned contained only incident reports, not near-miss or observation data. These reports tend to be more carefully filled out, using more formal English. It can also be postulated that it is easier for both those capturing the data and researchers labelling the datasets to identify precursor attributes in the case of an incident as there is less subjectivity in identifying key situational descriptors before a specific incident than in the case of unsafety.

Finally, the granularity of hyper-parameter optimisation in these papers was much finer than used during this investigation. This is because the grids of hyper-parameter values for these other investigations were selected with the aim of optimising the algorithm, rather than investigating the sensitivity and effects of changing the hyper-parameters.

Another comparison which could inform the use of automation to extract attributes is to compare the 'agreement' of the human annotation and the AI against the agreement between human annotators - which averaged at 49%. In this way, it is possible to assess whether the algorithm is performing on-par with human annotation. It is possible to calculate the 'agreement rate' in the same manner as in Table 5.19 using the obtained F1 score, as seen in Equation 5.7. Using the maximum F1 score achieved (54.2%), this research achieves an agreement of 37%. To achieve equal or better agreement than between human annotators, the model would need to obtain an F1 score of 66%.

$$Agreement = \frac{TP}{TP + FN + FP}$$

$$F1 = \frac{2TP}{2TP + FN + FP}$$

$$\frac{1}{Agreement} = \frac{TP + FN + FP}{TP} = 1 + \frac{FP + FN}{TP} = 1 + \frac{FP + FN + 2TP - 2TP}{TP}$$

$$= 1 + \frac{2(FP + FN + 2TP)}{2TP} - 2 = \frac{2}{F1} - 1 = \frac{2 - F1}{F1}$$

$$Agreement = \frac{F1}{2 - F1} \tag{5.7}$$

Effect of training data volume

Figures 5.8, 5.9 and 5.10 show the effect on performance metric on increasing the amount of training data available. These were calculated at default hyper-parameters with no oversampling on the test data, having trained the models on the entire training data set (2919 labelled safety event descriptions). As is expected, increasing the training data available increases the performance for all three metrics: F1, precision and recall. This increase is a steep linear with no flattening out towards 100% training data used which indicates that training data volume is a significant limitation to the performance of these models.

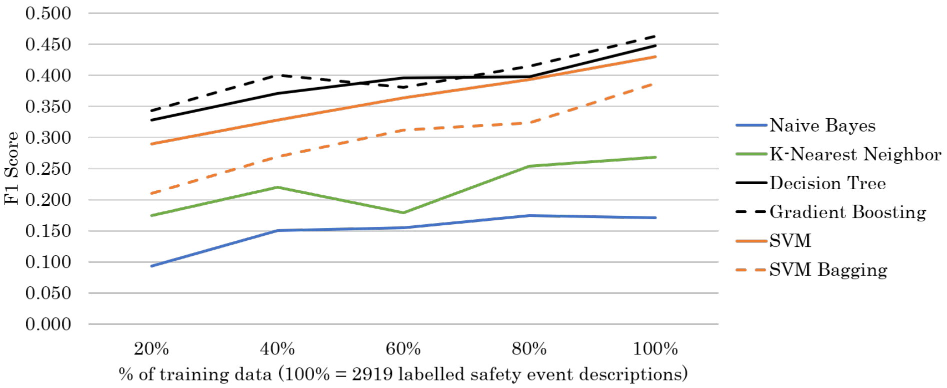


Figure 5.8: Line chart showing the effect of training data volume on F1 score

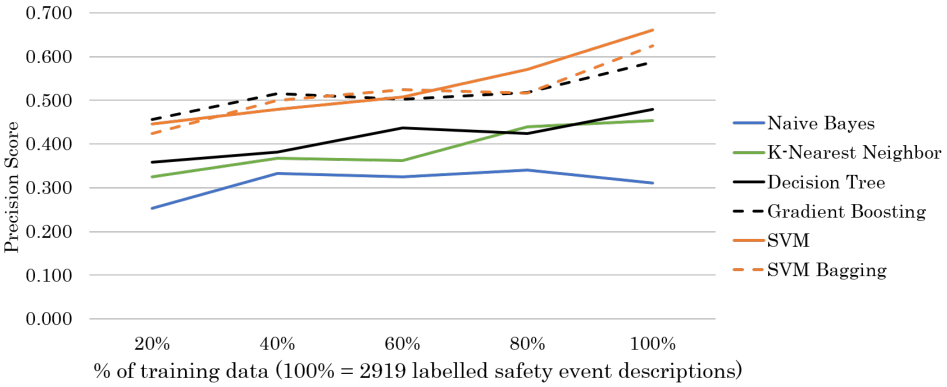


Figure 5.9: Line chart showing the effect of training data volume on Precision score

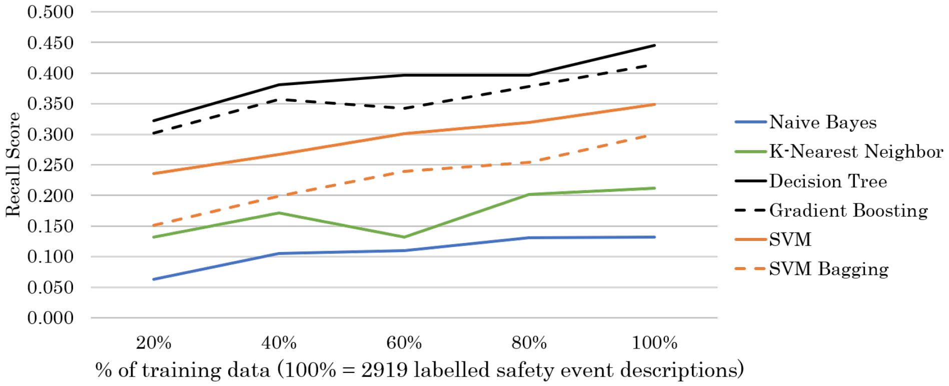


Figure 5.10: Line chart showing the effect of training data volume on Recall score

Dealing with class imbalance

Table 5.25 on p.143 demonstrates the effect of class imbalance on the model performance scores. Across all models, a greater proportion of positive examples facilitates better prediction, as seen by the increase in metric score towards the bottom of the table.

To mitigate against the class imbalance, deliberate oversampling of positive examples in the training set was used for training the algorithms. This oversampling was optimised with 10% granularity from minimum of 0% positive examples to 50% positive examples. Seen previously in Tables 5.23 and 5.24, optimising this oversampling for each attribute had a significant effect on the overall performance scores by increasing recall with minimal trade-off for precision. This relationship is also shown in Figure 5.11 for the final test data.

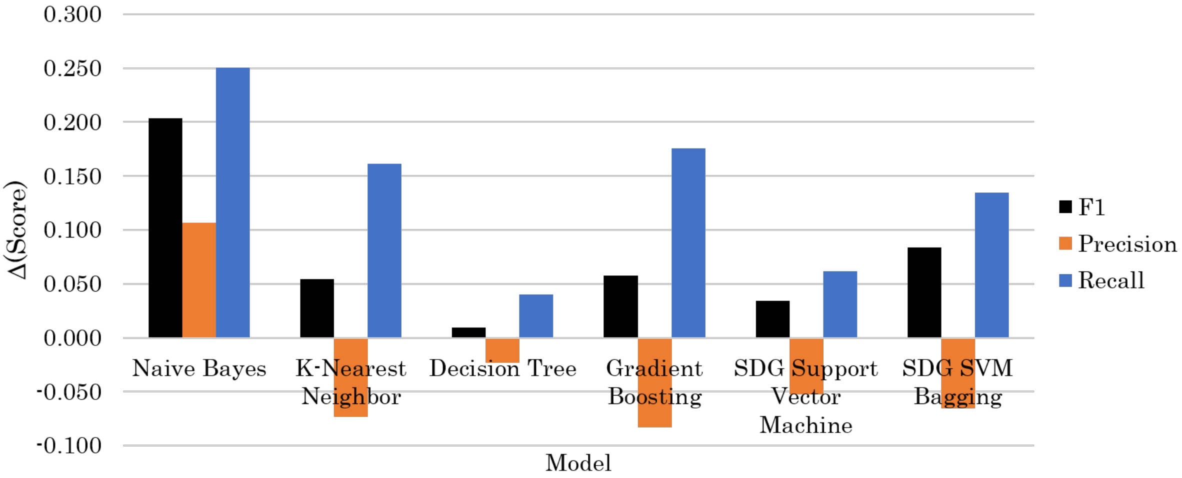


Figure 5.11: Bar chart showing the effect of optimising coarse-grid oversampling on the average performance metrics

Figure 5.12 shows the effect of the oversampling for individual attributes on the F1 score for three highest scoring methods: GB, SVM and SVM bagging. Each chart shows the value bins for change in metric score along the x-axis and the number of attributes which fell into that bin along the y-axis.

This is important as an average figure (such as in Figure 5.11) could be increased by a few attributes having an extremely large increase while the majority could have a slight decrease. If this were the case, it would make oversampling unsuitable for implementation. However, as can be seen, optimising the oversampling had a positive or neutral effect on the majority of F1 scores for attribute prediction.

Table 5.25: Attribute Prediction (default settings, no scale) by positive percent

		Naïve Bayes			k-Nearest Neighbour			Decision Tree		
Percentage	**No in bin**	**F1**	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**	**Prec**	**Recall**
(1, 1.5]	**14**	0.370	0.400	0.393	0.393	0.420	0.426	0.018	0.071	0.010
(1.5, 2]	**10**	0.293	0.310	0.296	0.274	0.308	0.266	0.114	0.181	0.092
(2, 2.5]	**4**	0.424	0.458	0.435	0.452	0.548	0.412	0.208	0.544	0.143
(2.5, 3]	**9**	0.409	0.406	0.435	0.434	0.491	0.408	0.047	0.230	0.027
(3, 3.5]	**1**	0.157	0.152	0.166	0.210	0.404	0.145	0.000	0.000	0.000
(3.5, 4]	**4**	0.461	0.493	0.444	0.472	0.580	0.435	0.147	0.378	0.093
(4, 4.45]	**4**	0.481	0.476	0.502	0.495	0.585	0.465	0.283	0.785	0.189
(4.5, 5]	**2**	0.443	0.439	0.459	0.444	0.500	0.438	0.084	0.583	0.047
(5, 10]	**8**	0.425	0.433	0.428	0.414	0.617	0.321	0.261	0.689	0.174
(10, 20]	**4**	0.484	0.486	0.484	0.497	0.725	0.383	0.463	0.713	0.360

		Gradient Boosting			Support Vector Machine			SVM Bagging		
Percentage	**No in Bin**	**F1**	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**	**Prec**	**Recall**
(1, 1.5]	**14**	0.300	0.564	0.230	0.351	0.600	0.284	0.230	0.450	0.175
(1.5, 2]	**10**	0.215	0.547	0.149	0.274	0.527	0.209	0.149	0.333	0.114
(2, 2.5]	**4**	0.478	0.815	0.352	0.505	0.682	0.412	0.410	0.632	0.337
(2.5, 3]	**9**	0.351	0.780	0.243	0.426	0.723	0.321	0.205	0.661	0.135
(3, 3.5]	**1**	0.221	0.650	0.138	0.248	0.444	0.183	0.183	0.567	0.111
(3.5, 4]	**4**	0.467	0.683	0.377	0.513	0.657	0.435	0.277	0.563	0.198
(4, 4.5]	**4**	0.531	0.738	0.433	0.574	0.712	0.496	0.451	0.706	0.356
(4.5, 5]	**2**	0.423	0.652	0.339	0.470	0.580	0.406	0.273	0.438	0.202
(5, 10]	**8**	0.420	0.728	0.305	0.468	0.645	0.378	0.315	0.616	0.222
(10, 20]	**4**	0.545	0.681	0.458	0.556	0.623	0.507	0.451	0.605	0.369

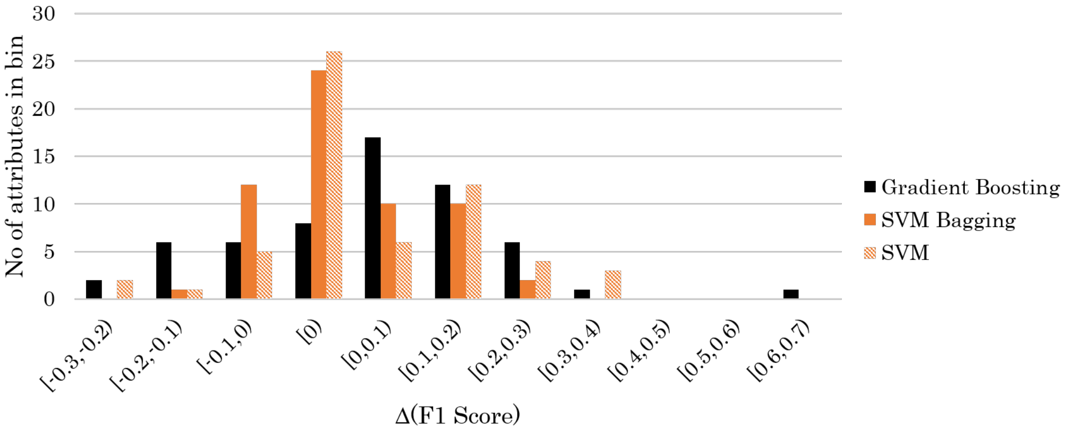


Figure 5.12: Bar chart showing the effect of oversampling on F1 score

Type of attribute

Figures 5.13, 5.14 and 5.15 illustrate the differences in the average performance scores for the three types of precursor attribute: actions, objects and site environment. Note that, as seen in Table 5.1, there are uneven numbers of attributes in each of these types. Specifically, for those occurring in over 1% of the data, there are 9 actions, 38 objects and 13 site environment descriptors.

These charts show that recall for site environment attributes outperforms those of actions and objects, however there is no significant differences in model performance between types of attributes. This means that a single, best performing algorithm can assess for all types, rather than selecting a different model for each attribute type.

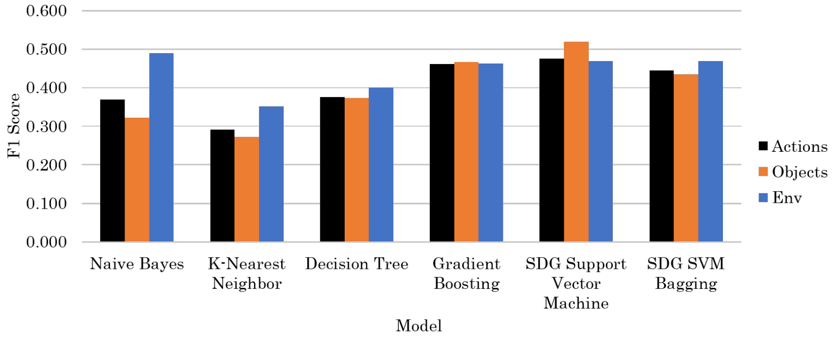


Figure 5.13: Bar chart showing the effect of attribute type on F1 score

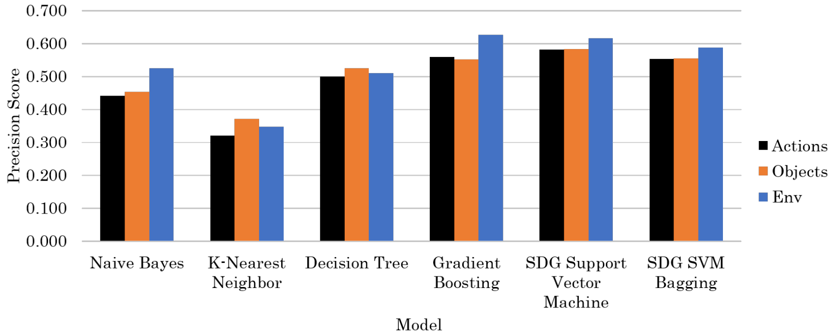


Figure 5.14: Bar chart showing the effect of attribute type on Precision score

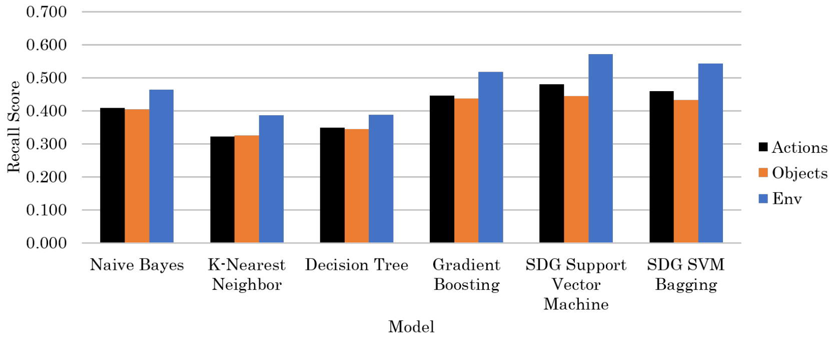


Figure 5.15: Bar chart showing the effect of attribute type on Recall score

Hyper-parameter optimisation

As previously mentioned, hyper-parameter optimisation is a key step in implementing any ML model and can have significant impacts on the performance of the algorithm. Each model algorithms was coarsely optimised with selected hyper-parameters, as introduced in Tables 5.13 to 5.17. Figure 5.16 illustrates the average increase in performance metric for each model. This is significant across all models with the exception of decision tree.

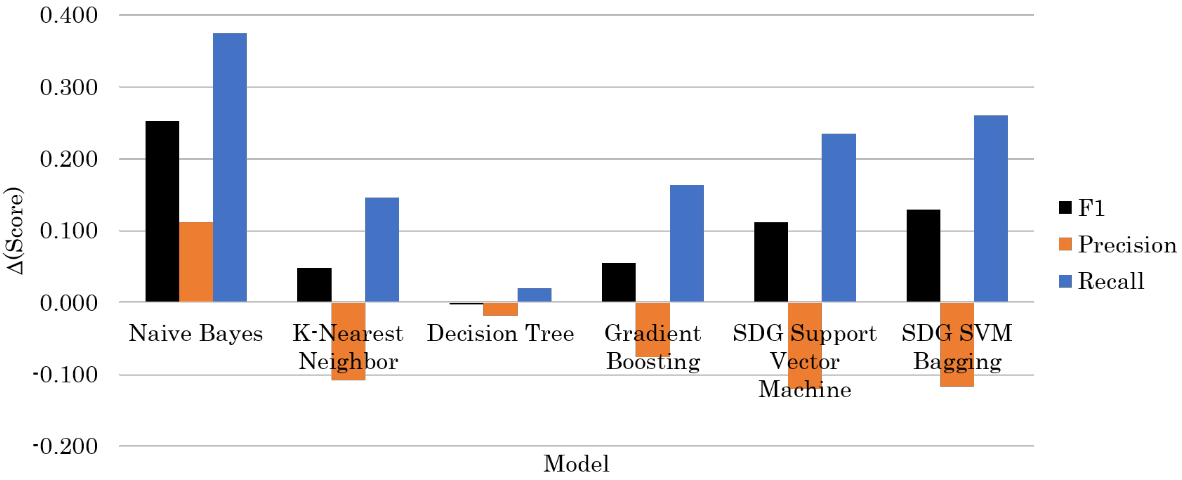


Figure 5.16: Bar chart showing the effect of coarse optimisation on the average performance metrics for 5-fold validation data

5.6 Summary

The qualitative analysis presented in Chapter 4 identified weaknesses in current learning from failure processes in the construction industry. one of which was an inability to systematically learning from small consequence failure events. Of the data collected from these events, the most pertinent information for learning is contained within the text fields documenting the event itself because of the high proportion of incomplete field entries and risk of subjectivity elsewhere in the forms.

Transforming the unstructured text descriptions of failure events, in this case safety incidents and observations, is essential to be able to use these data in digital analysis. This chapter presented 'attribute-based representations' as a structured representation of these text data. Previous research from University of Colorado developed a set of safety event attributes (Desvignes, 2014), which were used as the starting point in this analysis.

The worksite attributes fell under the following three categories: objects, actions and site environment descriptors. These attributes aim to objectively describe the worksite and can be identified prior to work commencing (or an incident occurring). This research identified 250 unique work attributes were identified in a set of 3,242 safety reports from a large UK construction company. However, only 60 attributes (listed in Table 5.21) occurred in 1% or more of the safety descriptions. These 60 attributes fully described 81% of the descriptions, i.e. 81% of annotated descriptions were not annotated with any of the less frequent attributes.

To manually label these text descriptions took many man hours. In order to implement attribute-base representation in industry, an automated method for identifying these attributes is necessary.

Text classification was selected as an appropriate ML task to automatically produce these attribute representations. This method was formed of two-parts: (1) converting the text data into a numerical vector, and (2) employing supervised text classification ML algorithms to predict the presence of these attributes.

Despite the increased semantic information and complexity of the word embedded representations, previous research has not yet shown this translates into significantly improved task metrics. Therefore, vector representations (i.e. 'Bag-of-words') were used for representing the text as a numerical vector. This method is explainable as it simply counts the words present and stores the information in a long, sparse vector. In this research, frequent bi-grams and stemming was also employed to mitigate the loss of semantic information.

Several classification algorithms were then investigated to predict the attribute classes from the numerical vector. The four algorithm types investigated were: Naïve Bayes, k-Nearest Neighbour, Decision Tree and SVM (Support Vector Machines). Two ensemble methods - gradient boosting for Decision Trees and bootstrap aggregating (Bagging) for SVM - were also applied to the final two algorithms respectively.

Overall, SVM (F1 = 54.2%) and Gradient Boosting (F1 = 51.8%) achieved the best performance scores. There was no advantage to implementing SVM bagging. Due to it added complexity, being an ensemble algorithm, Gradient Boosting is less explainable, and it is also slightly lower performing than SVM. Therefore, SVM is used to predict the attributes for the analysis in Chapter 6.

Further findings extracted from the exploration of these results are:

1. The relative small volume of training data inhibits the potential of all methods investigated. More labelled training data would probable significantly increase the performance metrics.
2. Attributes with a higher number of positive examples, i.e. those which occur more frequently in the data, achieve higher classification performances.

3. Oversampling had a significant positive effect on approximately half of attributes predicted due to the increase in recall outweighing the decrease in performance. Oversampling for these imbalanced data should be employed in the future.
4. The type of attribute - object, action or environment - had no significant effect on the classifiers' performances.
5. Hyper-parameter values, as expected, have a significant effect on the model performance. For use in industry, optimising - but not over-fitting - these will be key.

It should also be noted, however, that these classification algorithms are still only performing at 75% of human agreement scores. They would need to achieve 66% F1 to outperform human annotators, compared to SVM at F1= 54.2%. The implications of this are discussed in Chapter 7.

Chapter 6

Structured data to knowledge: Methods using structured attribute sets for knowledge discovery

"*Success is Foreseeing Failure*"

'To Engineer is Human' by Henry Petroski

Foreseeing (and therefore preventing) failure appears to be the goal for much of the analysis on past failure events to date. This chapter is included as I believed it to be important to demonstrate how structuring text data as a set of fundamental attributes can facilitate many different further analysis tasks. Here, I explore a method which could be applied to 'foresee' failure, in the form of risk analysis, but also explore methods which facilitate sense-making of the collective failure set. These explorations of different analysis methods help frame and guide discussion in Chapter 7 on how these technical processes can harmonise with social constructs on construction projects to create systematic organisational learning processes.

6.1 Knowledge Discovery: Data to Information

Successfully structuring free-text descriptions of events as a series of key attributes is an interesting and non-trivial task; however, alone, this achievement is not useful to a construction professional. The useful information required to gain value from these attributes comes from obtaining and analysing the patterns and relationships they have to each other and the event outcome. Presented here are three of many knowledge discovery methods to obtain useful information from the structured data. These methods are presented as illustrative examples only, not as complete exemplars of the methods, in order to facilitate discussion in the next chapter.

While the methods used to obtain patterns and relationships in data are often referred to as knowledge discovery methods, as discussed in Chapter 2, Section 2.4, the 'knowledge' of 'knowledge discovery' does not refer to the specific level of Tuomi's hierarchy. It would be more accurate to refer to these tasks as 'information generating'.

The three example methods presented here are:

- Use of attributes to calculate quantitative risk
- Use of attributes to predict incident outcomes
- Use of attributes for network analysis

6.1.1 Two-tail paired sample t-test

In this chapter, attribute labelling (using the SVM algorithm as detailed in the previous chapter) has been performed on the entire data set, both labelled and unlabelled, amounting to 14,882 safety events. As this includes the unlabelled data, it is necessary to have some measure of how well the attributes for these documents are being labelled. For this, a two-tail paired sample t-test is performed using the attribute frequencies, comparing the manually labelled data to the predicted labels.

It is assumed that the labelled data is representative of the entire data and therefore the proportions of documents relating to each attribute should be within significance range. Pair-sample t-tests pose the null hypothesis that the pairwise difference between the two samples is equal, i.e. $H0 : \mu(d) = 0$.

In comparing the labelled attribute frequencies to the attribute frequencies predicted for the entire data set, for the set of 60 attributes occurring $> 1\%$, the obtained p-value is 0.953%. This means the null hypothesis is accepted, with significant confidence (98.1%).

For this analysis, accepting the null hypothesis can be interpreted as signifying that the predicted attributes are representative of the labelled data.

6.2 Method 1: Use of attributes to quantify activity risk

6.2.1 Method

In this section, the attributes are used to calculate quantitative risk data for site tasks. By using the observed frequencies of attributes and event outcomes, it is possible to calculate the relative risk of an outcome quantitatively with finer granularity than the existing approaches, such as by trade, for example by Sooyoung and Fernanda (2016).

This approach applies Bayes Theorem (Equation 6.1). In the example presented only events which resulted in an accident in the following five sub-categories are included: *Fall from height*, *Handling, Lifting or Carrying*, *Hit/Struck by moving or falling object*, *Hit/Struck by something fixed or stationary* or *Slip, Trip or Fall on same level.* By calculating $P(A_n|Pr)$ (the probability an accident of a specific sub-category occurs given a precursor attribute is present), construction professionals can focus risk assessments and safety efforts on high scoring accident categories for their tasks.

$$P(A_n|Pr) = \frac{P(Pr|A_n) \times P(A_n)}{P(Pr)} \tag{6.1}$$

In Equation 6.1, $P(Pr|A_n)$ is the observed probability of an attribute given the accident sub-category, $P(A_n)$ is the observed probability of an accident of a specific sub-category and $P(Pr)$ is the observed probability of a precursor attribute *given that an accident has occurred*, $P(Pr) = P(Precursor_Attribute|Accident)$.

It is important to note that the data collected from incidents is unable to directly give an absolute probability of a precursor attribute or accident sub-category on project, only the probability given that a safety event has occurred.

For accident sub-category, the observed AFR (accident frequency rate) can be used to obtain this absolute probability, as seen in Equation 6.2 (where $P(A) = AFR =$ probability of accident per 10,000 man hours and it is given that $P(A|A_n) = 1$, i.e. if an accident of a specific sub-category has occurred an accident has certainly occurred). In calculating these results, an industry average AFR was used. In reality, the individual project AFR could be used, however, the data was not aggregated in this way. Nevertheless, in presenting a methodology, use of the industry AFR is acceptable.

$$P(A_n) = \frac{P(A_n|A) \times P(A)}{P(A|A_n)} = \frac{P(A_n|A) \times AFR}{1} \tag{6.2}$$

However, for probability of a precursor attribute $P(Pr)$, there is currently no equivalent data in order to calculate the absolute probability of a precursor attribute occurring on project. It is therefore necessary to bring this term to the left hand side of the equation. For tasks where the presence of an attribute is given, for example during preparation of a risk assessment for a specific task, $P(Pr) = 1$ and therefore the most likely accident categories can be identified and preventative measures employed. However, for comparison of the 'riskiness' of each attribute, each result is scaled by the frequency of the attribute on project (could also be described as the total exposure to the attribute) resulting in a skewed view. This limitation is discussed further after the example results are presented.

$$P(A_n|Pr) \times P(Pr) = P(Pr|A_n) \times P(A_n|A) \times AFR \tag{6.3}$$

6.2.2 Example results

The example results presented quantify risk given the attribute for nine frequently occurring attributes: four actions (Driving, Exiting/entering, Lifting/pulling/manipulating, Walking/moving around), three objects (Cabin, Machinery, Unpowered hand tool), and two site environment descriptors (Uneven surface, Slippery surface). It was chosen to present these attributes as they have the most demonstrable results to facilitate discussion of this method.

The results for the entire data - consisting of 4149 accidents from 152 project - are presented as well as two project case studies - both large highways projects. Project 1 data consists of 131 accident reports and Project 2 consists of 106 accident reports.

Table 6.1 and Table 6.2 show the result for the risk of a given accident sub-category given that the attribute on the right is present ($P(A_t|Pr)$). They also contain a row for the observed probability of an accident sub-category given that an accident occurred ($P(A_t|A)$) and a column for the observed probability of an accident sub-category given that an accident occurred ($P(Pr|A)$). Those cells shaded a deeper red highlight the higher relative risk.

For the most part, the results given are unsurprising. For example, '*Handling, Lifting or Carrying*' injuries are at higher risk from the action 'lifting/pulling/manipulating' while '*Slip, Trip or Fall on same level*' injuries are at higher risk from the action 'walking/moving around' and the two site environments of 'Uneven surface' and 'Slippery surface'.

The presence of these logical results adds face validity to the method. Additionally, these results are useful when planning a new activity as they provide empirical data to support anecdotal knowledge in order to provide evidence to focus safety efforts. However, there are two limitations of this:

1. As previously mentioned, this method only gives risk **given that** a precursor attribute is present (i.e. $P(Pr) = 1$). This means that frequent actions, objects or site environments will be over-represented in the risk comparisons. For example, the action 'walking/moving around' is seen to be the second highest risk attribute.
2. All injuries are considered equal. In this analysis, an injury which results in life-changing injury contributes equally to the risk than a sprained ankle. This is clearly unrepresentative.

Table 6.1: Table showing risk for all data - 4149 accidents

	Attributes	**Fall from height**	**Handling, Lifting or Carrying**	**Hit/Struck by moving or falling object**	**Hit/Struck by something fixed or stationary**	**Slip, Trip or Fall on same level**	
	$P(A_t\|A)$	2.7E-02	3.1E-01	1.6E-01	1.0E-01	1.5E-01	$P(Pr\|A)$
Actions	**Driving**	1.2E-06	1.0E-05	1.8E-05	5.8E-06	4.6E-06	3.6E-02
	Exiting/entering	1.6E-05	4.2E-05	3.2E-05	2.9E-05	6.8E-05	9.1E-02
	Lifting/manipulating	2.0E-05	5.1E-04	2.5E-04	1.4E-04	1.0E-04	5.5E-01
	Walking	4.0E-05	1.7E-04	5.3E-05	5.8E-05	2.9E-04	2.9E-01
Objects	**Cabin**	1.7E-06	1.7E-05	1.4E-05	9.8E-06	2.8E-05	3.8E-02
	Machinery	5.8E-06	3.6E-05	3.2E-05	9.3E-06	1.3E-05	5.4E-02
	Unpowered tool	0.0E+00	4.9E-05	4.1E-05	1.7E-05	4.0E-06	6.7E-02
Site	**Slippery surface**	1.0E-05	7.6E-05	1.4E-05	7.5E-06	9.7E-05	9.4E-02
	Uneven surface	3.6E-05	1.5E-04	2.9E-05	1.3E-05	2.6E-04	2.2E-01

Table 6.2: Table showing risk for Project 1 data - 131 accidents

	Attributes	**Fall from height**	**Handling, Lifting or Carrying**	**Hit/Struck by moving or falling object**	**Hit/Struck by something fixed or stationary**	**Slip, Trip or Fall on same level**	
	$P(A_t\|A)$	1.5E-02	2.1E-01	1.8E-01	3.8E-02	2.4E-01	$P(Pr\|A)$
Actions	**Driving**	0.0E+00	1.8E-05	3.7E-05	1.8E-05	1.8E-05	8.4E-02
	Exiting/entering	0.0E+00	0.0E+00	9.2E-05	0.0E+00	7.3E-05	1.1E-01
	Lifting/manipulating	1.8E-05	3.7E-05	5.5E-05	1.8E-05	4.2E-04	5.5E-01
	Walking	1.8E-05	3.7E-05	1.8E-05	0.0E+00	4.0E-04	2.6E-01
Objects	**Cabin**	0.0E+00	1.8E-05	0.0E+00	1.8E-05	1.8E-05	9.9E-02
	Machinery	0.0E+00	5.5E-05	1.1E-04	0.0E+00	0.0E+00	1.2E-01
	Unpowered tool	1.8E-05	0.0E+00	9.2E-05	0.0E+00	7.3E-05	1.2E-01
Site	**Slippery surface**	0.0E+00	5.5E-05	5.5E-05	5.5E-05	7.3E-05	1.3E-01
	Uneven surface	0.0E+00	0.0E+00	1.8E-05	0.0E+00	2.7E-04	2.4E-01

6.2.3 Mini-discussion: what do these results mean?

In his undergraduate dissertation, using the labelled data from this project, Campbell (2020) applied an injury severity score (ISS), inspired by Hallowell and Gambatese (2009b), to adjust for the second limitation and an artificially constructed attribute exposure data set to demonstrate the effect of the first. The further limitations here were that the ISS is subjective, ranking injury types on a score basis which had some medical foundation but failed to capture the long-term effects of the injury or the nuances between injuries of a similar type. Also, as he stated, his use of an artificial exposure data set means that the results can only be used to demonstrate the effects of attribute frequency and cannot be used to form any conclusions.

In reality, obtaining accurate exposure data for the attributes will be tricky. Some information may be available via planning schedules and site diaries, but consistently achieving the granularity of information required will be non-trivial. This limits the usefulness of a quantitative, probabilistic method as direct comparison between the attributes is distorted by their frequency.

Campbell, 2020's dissertation also explored the likelihood of different injury types (e.g. laceration, sprain etc) rather than accident sub-category. This demonstrates that this method can also be applied to different outcome varieties, not just the set of classes presented here.

Comparison of the score achieved to a baseline or between projects may be a far more immediately promising method. Table 6.3 illustrates the difference in risk scores for Project 1 vs the risk scores for the full data set. This clearly highlights in green those areas which Project 1 is outperforming the baseline it is being compared to (i.e. the risk of an accident type is lower) and areas which should be targeted for improvement (i.e. the risk of an accident is higher, shown in red). Here, it can be seen that Project 1 had a significant higher proportion of their accidents recorded as '*Slip, Trip or Fall on same level*' especially those involving 'lifting/manipulating'. On the other hand, accidents resulting from '*Handling, Lifting or Carrying*' appear to be below the baseline as do accidents involving 'walking/moving around'. For decision-makers on projects, this information can be used to target particular activities.

However, different project types have different activity frequencies so comparing the results in this way, against other infrastructure projects in multiple sectors, is insufficient to identify true performance differences. By comparing the risk score of one project against a similar project type, the different values are more representative. Table 6.4 illustrates the difference in risk scores for Project 1 vs Project 2, both large highways projects. This confirms the result that '*Handling, Lifting or Carrying*' injuries have a reduced risk on Project 1, however, '*Slip, Trip or Fall on same level*' injuries are on a comparable level except for the higher value involving 'lifting/manipulating'. For Project 1, this could indicate that their workers are attempting awkward manoeuvres and tripping and instigate an investigation into underlying causes. Additionally, Project 1 and 2 could also compare their processes in light of these results to understand where the other is doing things well.

Despite the limitations, the results obtained by this method are good at identifying relative risk given that the attribute is present as well as highlighting similar projects doing things well or under-performing. Discussed further in Chapter 7, this method can be used for a variety of tasks, such as dis-aggregated safety reporting and to enhance risk assessments on project sites. Additionally, the information gained via this method can be used to compare risk profile of projects from the design stage, integrating with BIM technology, to inform design choices.

Table 6.3: Table showing difference in risk for Project 1 compared to all data

	Attributes	**Fall from height**	**Handling, Lifting or Carrying**	**Hit/Struck by moving or falling object**	**Hit/Struck by something fixed or stationary**	**Slip, Trip or Fall on same level**	
	$P(A_t\|A)$	-1.1E-02	-9.9E-02	2.6E-02	-6.5E-02	8.8E-02	$P(Pr\|A)$
Actions	**Driving**	-1.2E-06	7.9E-06	1.9E-05	1.3E-05	1.4E-05	4.8E-02
	Exiting/entering	-1.6E-05	-4.2E-05	5.9E-05	-2.9E-05	5.6E-06	1.6E-02
	Lifting/manipulating	-1.9E-06	-4.7E-04	-1.9E-04	-1.2E-04	3.2E-04	2.3E-03
	Walking	-2.2E-05	-1.3E-04	-3.5E-05	-5.8E-05	1.1E-04	-3.2E-02
Objects	**Cabin**	-1.7E-06	9.7E-07	-1.4E-05	8.5E-06	-9.4E-06	6.1E-02
	Machinery	-5.8E-06	1.9E-05	7.8E-05	-9.3E-06	-1.3E-05	6.8E-02
	Unpowered tool	1.8E-05	-4.9E-05	5.1E-05	-1.7E-05	6.9E-05	5.5E-02
Site	**Slippery surface**	-1.0E-05	-2.1E-05	4.1E-05	4.7E-05	-2.4E-05	3.6E-02
	Uneven surface	-3.6E-05	-1.5E-04	-1.1E-05	-1.3E-05	1.9E-05	1.2E-02

Table 6.4: Table showing difference in risk for Project 1 compared to Project 2

	Attributes	**Fall from height**	**Handling, Lifting or Carrying**	**Hit/Struck by moving or falling object**	**Hit/Struck by something fixed or stationary**	**Slip, Trip or Fall on same level**	
	$P(A_t\|A)$	5.8E-03	-1.5E-01	7.0E-02	-9.0E-03	-8.6E-03	$P(Pr\|A)$
Actions	**Driving**	0.0E+00	-4.3E-06	1.4E-05	1.8E-05	-4.3E-06	8.5E-03
	Exiting/entering	0.0E+00	-6.8E-05	6.9E-05	-2.3E-05	-1.7E-05	2.2E-02
	Lifting/manipulating	-4.3E-06	-6.9E-04	-1.3E-04	-2.7E-05	2.2E-04	-1.6E-02
	Walking	1.8E-05	-3.1E-05	-2.7E-05	0.0E+00	-1.4E-04	-3.3E-02
Objects	**Cabin**	0.0E+00	1.8E-05	-2.3E-05	-4.3E-06	-1.2E-04	1.4E-02
	Machinery	0.0E+00	-5.8E-05	4.2E-05	0.0E+00	-6.8E-05	-5.0E-04
	Unpowered tool	1.8E-05	-9.1E-05	6.9E-05	0.0E+00	7.3E-05	5.6E-02
Site	**Slippery surface**	0.0E+00	5.5E-05	5.5E-05	5.5E-05	-8.5E-05	6.4E-02
	Uneven surface	0.0E+00	-1.4E-04	-2.7E-05	0.0E+00	-2.5E-04	-7.5E-02

6.3 Method 2: Use of predict incident outcomes

6.3.1 Method

This method demonstrates a classification task using the text descriptions of the safety events, structured using the attributes from the previous chapter, as an input to predict the outcome of the safety event. This is equivalent to the task presented in Baker, Hallowell, and Tixier (2020a) where the outcomes predicted were injury severity, injury type, body part impacted, and incident type. Here, just incident type is predicted to illustrate the method using the UK infrastructure data and attributes from this analysis.

As in Baker, Hallowell, and Tixier (2020a), the accident outcome is independent from the attribute extraction task. This outcome was taken from the original data collected by the personnel on site (described in Section 5.2) as the 'Incident Sub Category' for accidents. Only sub-categories >20 positive occurrences were included for prediction. Those data associated with sub-categories occurring less frequently were removed from the training data set.

ML task and algorithms

The classification task performed here is a multi-class task, where each input is classified into one of several classes, unlike in the previous chapter where the task is binary classification.

For the machine learning task, three algorithms were investigated: SDG SVM, Gradient Boosting (both used previously) and Extreme gradient boosting (XGBoost). XGBoost (Chen and Guestrin, 2016) was included to have comparable results to Baker, Hallowell, and Tixier (2020a). This algorithm adds a regularization term to the loss function in order to penalize the complexity of the model and implements a number of optimization tricks to speed-up training.

As the purpose of this section was to illustrate the method, no algorithm optimisation was undertaken. Additionally, no over-sampling or under-sampling was undertaken to adjust the class imbalances.

As no optimisation was attempted, a single, randomised 30:70 test:train split was used. I decided to use a larger proportion of test data here as the complete data set is larger and none needs to be set aside for optimisation. Using a larger test set is desirable as it gives a better chance of getting a representative sample of the full data; however, this leaves less data for training.

Model Performance Assessment

For assessing the performance of multi-class classification, confusion matrices can be constructed showing the class predicted vs the class expected (as shown previously for binary classification). For the algorithms here, a recall, precision and F1 score is obtained for each class. These can then be averaged to give an overall score for the algorithm (macro average). As the classes are imbalanced, a weighted average can also be obtained - this is a better representation of the performance of the model.

Unlike in imbalanced binary classification, the overall accuracy can also be used to assess the performance of the model.

6.3.2 Example results

As seen in Table 6.5, Gradient Boosting marginally outperforms XGBoost. This was also found for text classification tasks in an online comparison of different boosting algorithms by Gursky (2020). However, these XGBoost results can be considered on-par with Baker, Hallowell, and Tixier (2020a), who achieved 42% average accuracy for accident sub-category outcome prediction, where the imbalance of the classes had been smoothed via re-sampling. The weighted average results of this analysis at 44% can be considered comparable considering no optimisation or re-sampling was performed.

However, predictive analysis using purely attribute-based inputs, both here and in published literature, have not achieved performance scores sufficient for implementing these methods in industry. Therefore, while this exercise demonstrates the applicability of the attributes, it also hints that work-site attributes do not tell the entire story. Work such as Alruqi and Hallowell (2019) state that work site attributes are not the only predictive features of a safety incident; we would expect an increase in prediction performance if other features are included.

Table 6.5: Sub-category prediction for accident data

Accident Sub-category	**SGD SVM**			**XGBoost**			**Gradient Boosting**			**Number of**
	Prec	**Recall**	**F1**	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**	**examples**
Contact with moving machinery	0.00	0.00	0.00	0.56	0.26	0.35	0.52	0.31	0.39	35
Driving at work	0.21	0.38	0.27	0.30	0.38	0.33	0.30	0.38	0.33	8
Exposure to harmful substances	0.24	0.36	0.29	0.27	0.14	0.18	0.40	0.27	0.32	22
Fall from height	0.60	0.10	0.17	0.00	0.00	0.00	0.33	0.10	0.15	31
Falling dust / debris into eye	0.86	0.73	0.79	0.85	0.76	0.80	0.81	0.75	0.78	51
Handling, Lifting or Carrying	0.46	0.69	0.55	0.43	0.68	0.53	0.45	0.67	0.54	357
Hit/Struck by moving or falling object	0.43	0.17	0.25	0.49	0.24	0.32	0.43	0.25	0.32	208
Hit/Struck by something fixed or stationary	0.47	0.18	0.26	0.40	0.16	0.23	0.40	0.19	0.26	131
Injured by person or animal/insect	0.62	0.45	0.53	0.62	0.45	0.53	0.67	0.55	0.60	11
Slip, Trip or Fall on same level	0.50	0.72	0.59	0.55	0.68	0.60	0.57	0.65	0.61	194
Struck by moving vehicle	0.08	0.09	0.09	0.08	0.09	0.09	0.23	0.27	0.25	11
performance		**0.47**			**0.47**			**0.48**		
Macro avg	0.41	0.35	0.34	0.41	0.35	0.36	0.47	0.40	**0.41**	
Weighted avg	0.46	0.47	0.43	0.47	0.47	0.44	0.48	0.48	**0.46**	

6.3.3 Mini-discussion: what do these results mean?

Predicting the outcome of a safety accident is probably not 'useful' of itself as predicting the outcome of an event in hindsight seems redundant - if the incident is recorded, we already know the outcome. However, by proving that the attributes are predictive of the outcome, this result demonstrates the attributes' relevance to the safety - or unsafety - of a situation. Therefore, more interesting for a construction professional is that further analysis can be carried out to identify the 'most predictive attributes' (as in Baker, Hallowell, and Tixier (2020a)).

As an example of possible further analysis, Figures 6.1 to 6.3 show the five most and five least predictive using coefficients from the SVM prediction. A large positive coefficient (relating to $|w|$ from Table 5.10 in previous chapter) means its presence is a significant predictor while a large negative means its absence is a significant predictor.

Some of these significant relationships are unsurprising. In Figure 6.1, the most significant predictor of a '*fall from height*' is 'insufficient edge/fall protection' indicating that this is a key attribute to this injury type. Meanwhile, as it has a high negative coefficient, the presence of a 'high fence' may be a significant predictor that a '*fall from height*' is unlikely. However, interpreting these 'negative' coefficients must be done with caution as, due to the multi-class nature of the classification task used, if an attribute is an extremely high predictor of another class, then its presence precludes it from the others. For example, in Baker, Hallowell, and Tixier (2020a), they found that the absence of their attribute 'improper PPE' was predictive of all classes except 'PPE' and 'rules' related injuries. While this could be because having the correct PPE is insignificant in the other classes ('access', 'dropped item', 'equipment/tools' and 'slip/trip/falls'), it is more rational that the attribute is simply so predictive of 'PPE' related injuries that its presence is indicative that this class is correct. In this way, an attribute can occur with a large negative coefficient in all classes other than the one it is most associated with, not because its absence is predictive of the class but because its presence is so predictive of a single class.

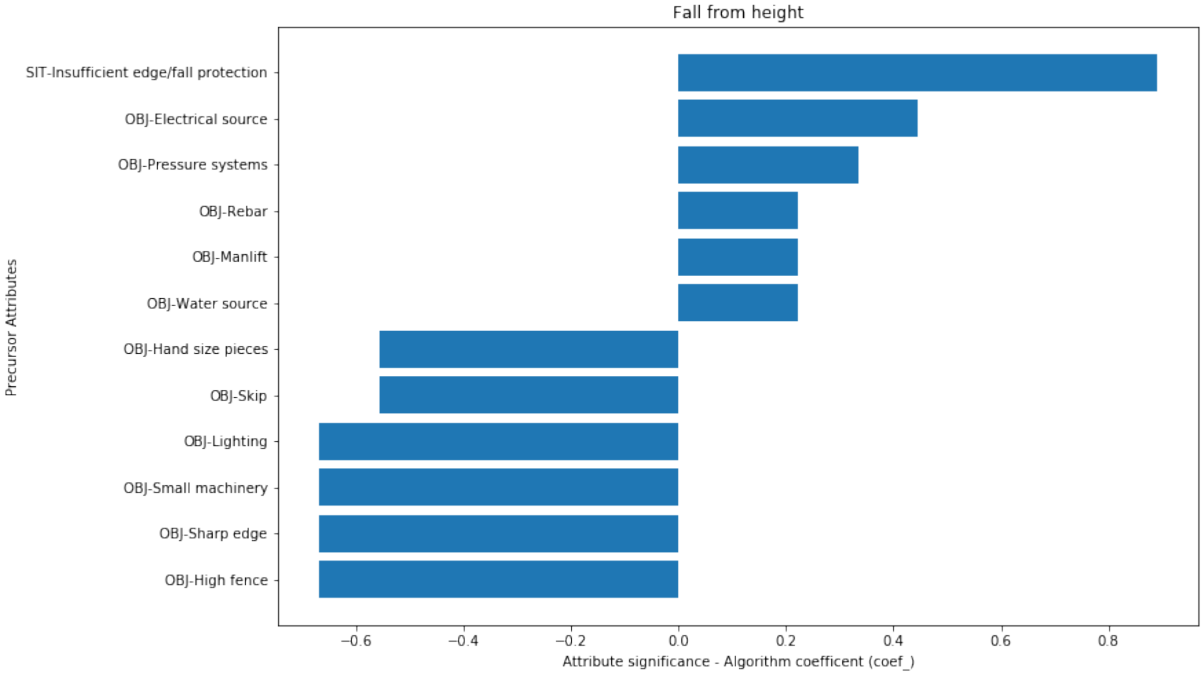


Figure 6.1: Attribute importance chart for accident sub-category '*Fall from height*'

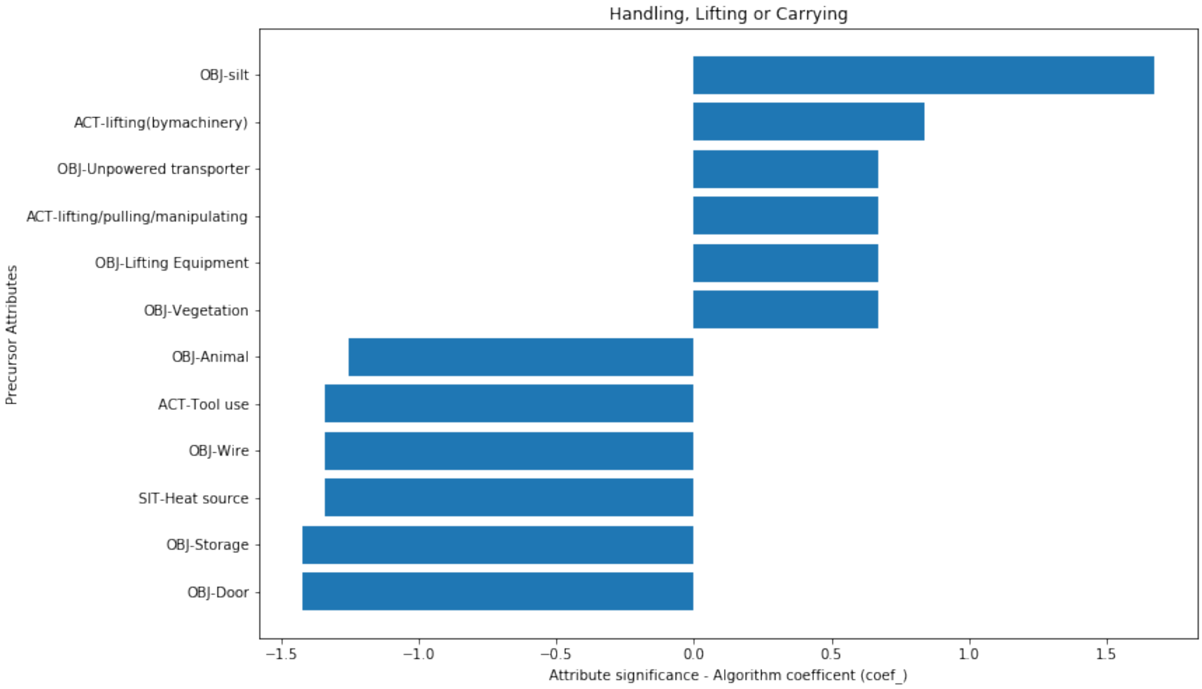


Figure 6.2: Attribute importance chart for accident sub-category '*Handling Lifting Carrying*'

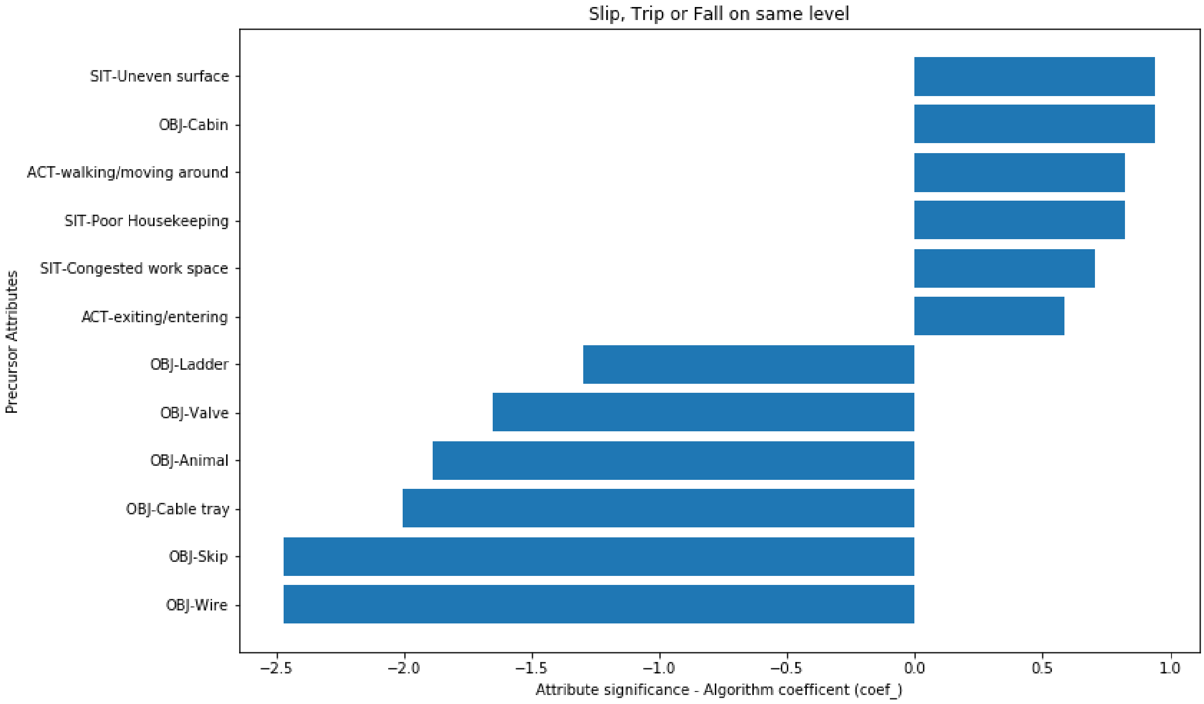


Figure 6.3: Attribute importance chart for accident sub-category '*Slip, trip, fall*'

6.4 Method 3: Use of attributes for network analysis

6.4.1 Method

Transforming the unstructured text descriptions into structured data in the form of attribute features allows network analysis methods to be performed.

Networks consist of nodes (variables) and the relationships between them (edges). In its most simplistic application, it can be used as a visualisation tool. This visualisation is what is presented here, where the relationship displayed is co-occurrence of the variables (attributes).

More complex networks can be built up by including outcome variables and passive variables like project sector, location, contract type etc. As well as including more sophisticated variables, the relationships (edges) can be more complex than simple co-occurrence. Borgatti and Ofem (2010) propose five different relationship types: interactions (such as co-occurrence used here), similarities, social relations (objective relationships such as sub-project of, joint venture partner), mental relations (subjective relationships such as 'likes', 'dislikes') and flows (of information, money, influence). Real value then lies in the further analysis of these complex networks using social network theory or systems analysis.

6.4.2 Example results

Figures 6.4 to 6.7 show four example networks of co-occurring attributes. In each pair of graphs, the manually labelled data are displayed in the first graph and the automatically labelled full data set in the second. The first pair show data which resulted in an injury, while the second pair are for near-miss/safety observation data.

In these network graphs, strong links are in darker, purple colours. Additionally, node size scales with attribute frequency. In each case the top 20 attributes were included. The minimum number of co-occurrences required to draw an edge varies to achieve a clear graph.

6.4.3 Mini-discussion: what do these results mean?

Accident Data

The first set of graphs, Figures 6.4 and 6.5, illustrate co-occurrence of attributes for accident data. The first observation to note is that the set of attributes for the manually and automatically labelled are extremely similar, with 17/20 exact matches, reinforcing the previous t-test calculation (sub-section 6.1.1).

A notable exception is the absence of 'light vehicle' in the top 20 of the automatically labelled data which was a significant sized node in the labelled data. However, this does not necessarily mean that the automatic labelling is not representing this correctly. As the labelled data had a large number of rail and highways projects included, it could be that these projects have a larger number of incidents and observations involving light vehicles than the other sectors. Therefore, this would be a lower ranked node in the full data set.

There is a significant edge between 'walking/moving around' and 'uneven surface' in accident data showing that these two attributes co-occur frequently for both the manually labelled and automatically labelled data. These two attributes also have a strong connection to 'lifting/pulling/manipulating' suggesting that a combination of these three attributes is a sign of 'unsafety'.

This inclusion of the relationships between these attributes is one of the main advantages network analysis has over the other, numerical methods presented in this section. Even just using this method as a visualisation tool, it is clear which pairs or groups of attributes are problematic, rather than the occurrence of an attribute in isolation.

Near-miss data

In near-miss examples (Figures 6.6 and 6.7), there are also 17/20 attributes which are an exact match in the top 20 attributes for the manually and automatically labelled data. Again, this demonstrates the corroborates the t-test result.

Interestingly, in comparison to those events which resulted in an injury, there is far more emphasis on driving and various vehicle types in the near-miss data. This may indicate reporting bias rather than an actual increase. From my own observations on site, gate guards have more opportunity and means to report vehicular near-misses from their booths, than site workers do during their working time. On the other hand, this emphasis could be a representative view of the proportion of near-misses attributable to vehicular movement. To address the question of reporting bias, the name/position of the reporter could be recorded to add to the evidence base, however, this could have significant negative consequences on non-reporting and blame culture. This is further explored during the discussion on data collection for learning.

Nevertheless, these types of comparisons can provide construction professionals information about the types of activities being reported and the current ‘hot spots’ for safety as reported by their workers.

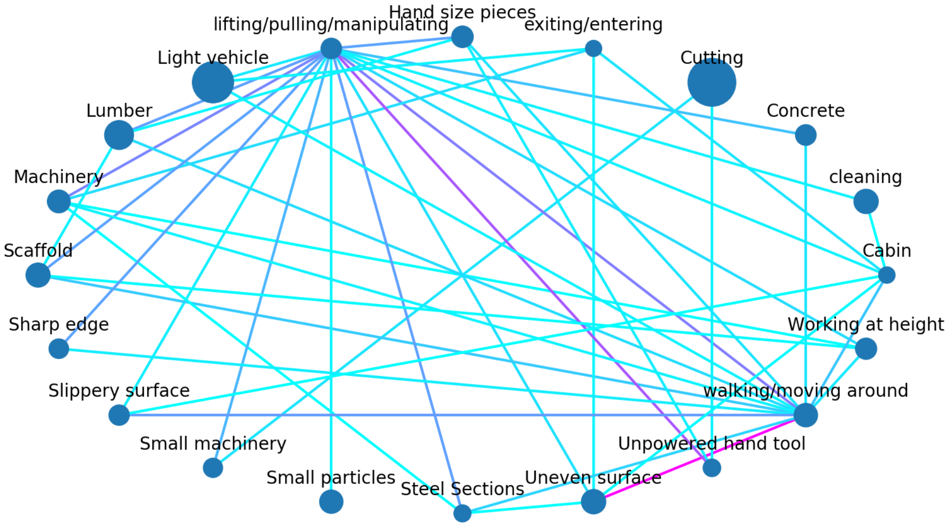


Figure 6.4: Network Graph for Labelled Accident Data of Attribute Co-occurrence (top 20 attributes which have over 5 co-occurrences)

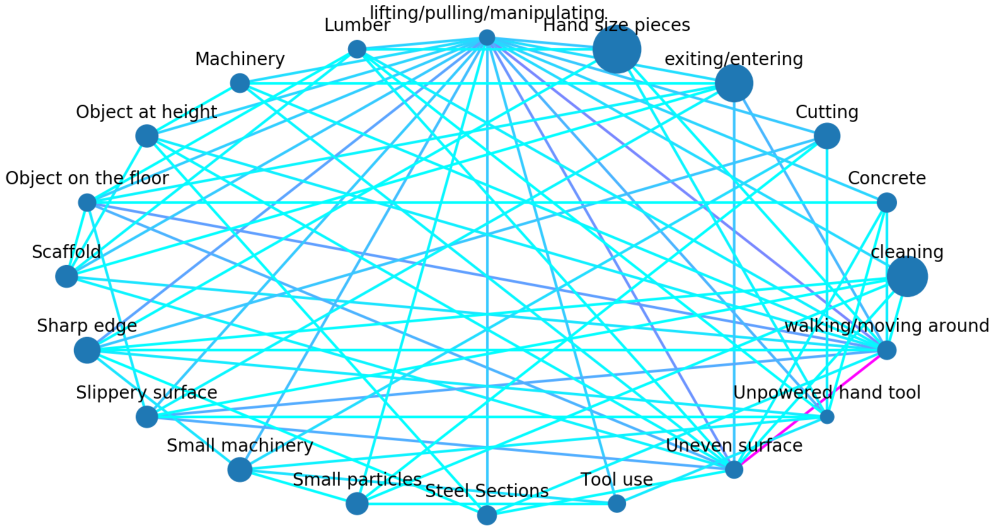


Figure 6.5: Network Graph for All Accident Data of Attribute Co-occurrence (top 20 attributes which have over 20 co-occurrences)

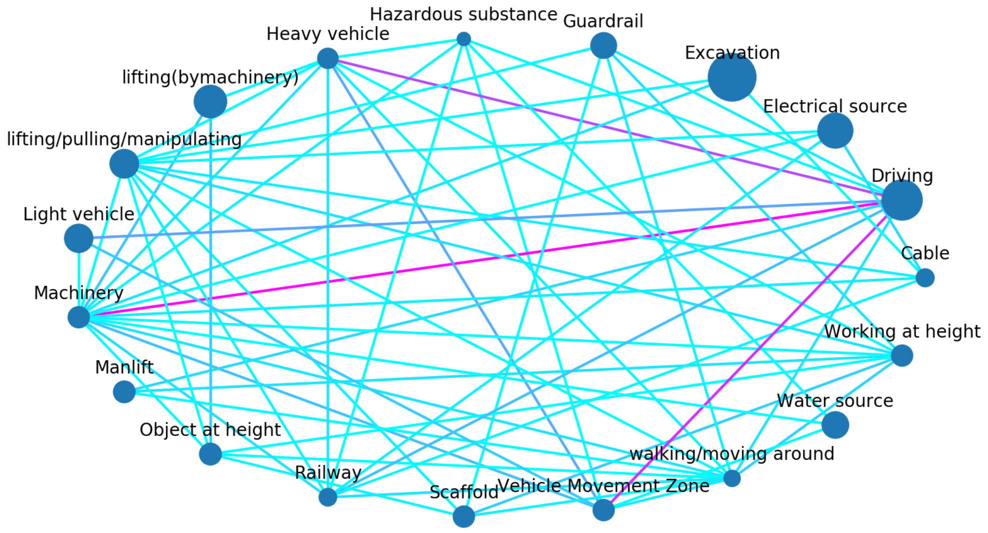


Figure 6.6: Network Graph for Labelled Near-miss Data of Attribute Co-occurrence (top 20 attributes which have over 10 co-occurrences)

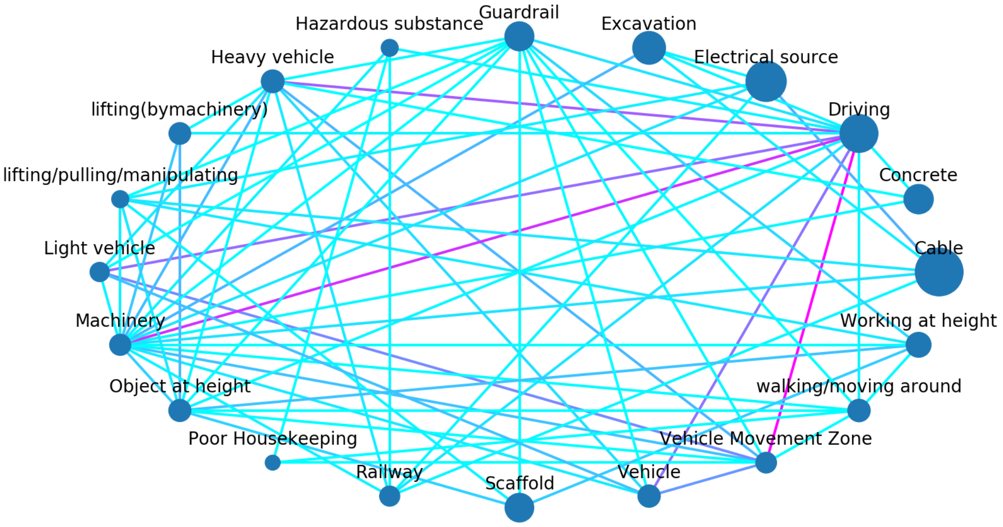


Figure 6.7: Network Graph for All Near-miss Data of Attribute Co-occurrence (top 20 attributes which have over 50 co-occurrences)

6.5 Summary

The three knowledge discovery methods presented in this Chapter demonstrate only a small range of the approaches unlocked by transforming unstructured text descriptions of failure events into a structured set of fundamental attributes. The information gained via application of these methods included forecasting future risk as well as sense-making of current/past failure collectives. These results could generate tremendous positive impact for organisational learning in the construction industry.

However, these methods are merely tools to enable a greater agenda: more efficient and systematic learning from failure in construction. It is, therefore, essential, having developed working examples of this technology, to examine these methods in light of this agenda and its context. This is covered next.

Chapter 7

Discussion

"*Truth emerges more readily from error than from confusion*"

by Francis Bacon in "Novum Organum" Bacon, 1869

This was my favourite chapter to write. In some ways, I found it cathartic and exciting to have space to explore and structure all the ideas and streams of thought concerning four years of research. Here, I address my final research question "How is best to implement this type of learning into systematic processes for the construction industry?" which develops my research back from the specific task of identifying attributes from text descriptions to the applicability of this methodology to learning from failure in the construction industry.

7.1 Does this research support application of attribute-based analysis for the construction industry?

Construction is complex. Distilling its complex nuances and contextualised information into a set of structured data facilitates quantitative, computer-based analysis. In this research, attributes have been presented as a method of structuring the text descriptions of failure events; however, it must be discussed whether it is suitable for the construction industry to treat this type of text data in this way. Therefore, this section examines whether the findings of both the semi-structured interviews in Chapter 4 and the results from the data labelling exercise in Chapter 5 support the use of attributes to structure the text data. This support is determined by examining whether attribute-based analysis, as developed here, stands up under scrutiny when considering the use cases presented, the context of the analysis and other data available.

7.1.1 The narrative so far

To initiate this discussion, three points previously presented should be re-stated. The first concerns the current "state-of-play" regarding learning from failure in the construction industry. The second point re-explores the issues with current data collection and those found during the analysis presented in Chapter 5 (subjectivity, number of fields etc). The third concerns the current rationale behind employing attribute-based frameworks to structure unstructured data.

Current "state-of-play"

Current learning from failure in construction is centred around human competence and industrial level change to standards from significant case-studies. As found in Chapter 4, these two factors have created a 'learning from failure' process which focuses on single-loop learning with responsibility lying at the individual level to implement change. In many cases, this is driven by dissemination of "alerts" following investigation of events with significant consequences, and reliance on individuals to have the time, access and inclination to pick these documents up and implement the lessons into their own work or projects.

For small consequence events, there simply isn't the stimulus to invest time and money into investigative efforts to analyse the event information and draw out what change is required. If this investigation did take place past a superficial conjecture at a root cause category, there would likely still be significant opposition, both culturally and financially, to implement change based on a single small, insignificant event. Despite the fact that when these smaller events co-occur, we observe catastrophic events as illustrated by the Swiss Cheese model (holes lining up in the cheese).

These factors lead to a requirement to more effectively use the data currently collected on failure events. Interviewees and my own experience confirmed that failures where the 'point of failure' is on-site - i.e. safety, quality, environment - have extremely similar data collection methods and processes. Of these, interviewees identified safety failures as the most significant form of failure and had the highest confidence in the completeness and uniformity of the data (see Section 4.3 in Chapter 4). Therefore, safety data was adopted for the remainder of the research. It should be made clear in this discussion how the findings here also relate to learning from other failure modes, and where the findings may not apply.

Limitations with current data

In observing the form of the data currently collected about safety failures, several key limitations must be addressed. The first is the number of manual fields, including a number of non-compulsory ones which have extremely low completion rates - especially for non-compulsory free-text fields. This includes 'details of action taken' (6%), 'investigation summary' (0.4%) and 'suggested improvements to senior management' (3%).

The second major limitation is the subjectivity of the categorical data fields. During the data labelling exercise in Section 5.5.1, it was found that annotators disagreed with 16.6% of immediate cause categories - 'Incident Sub-Category' - such as 'falling material' or 'contact with electricity'. This subjectivity is compounded when considering the biases (conscious or unconscious) which occur when relying on individuals to self-report their own, or their colleagues, failure events. The 4% of near-miss reports containing description of an injury, but not listed as an accident, are testament to the inadequacy of these multiple-choice categories.

Rationale behind attribute-based analysis

To address the inadequacies of the current structured data fields, both categorical and numerical, used for data analysis, previous research has set a precedent for use of key event attributes for construction safety, for example Desvignes (2014). The development of such attributes stemmed from the desire to quantify activity risks (see Esmaeili, 2012) with finer granularity than existing models, for example trade-based assessment. Pre-job risk identification has been proven to be effective in the prevention of safety mistakes. Does similar work exist for quality? In this research, two further advantages to attribute-based analysis have been hypothesised: anonymisation of the raw data and explainability of the result.

However, regardless of the theorised benefits, it needs to be considered whether the experience of attribute-based analysis for this research supports the further development of this methodological choice.

In Chapter 5, the annotation task resulted in a set of 250 unique attributes, of which only 60 occurred in more than 1% of the safety event descriptions. Despite this high proportion of infrequent attributes, 81% of the descriptions were fully described using the set of 60 attributes. By which, this means 81% of text descriptions were not labelled with any of the less frequent attributes. Additionally, 113/250 unique attributes occurred only once in the annotated data set.

The proportion of attributes (24%) which fully describe the majority of descriptions (81%) corroborates the "Pareto Prinicple" or 80:20 rule. Sanders (1987) notes that this principle, developed from the observations of engineering quality pioneer Joseph Juran, allows for prioritisation of action as, for many phenomena, 20% of variables will account for 80% of the results.

Two discussion points are raised by this high proportion of extremely infrequent attributes: the granularity of attributes needed to be representative of the text information and the granularity of attributes needed to be useful for analysis.

7.1.2 How representative are the attributes refined through this research?

A 'representation' of something (in this case a text description which is itself a representation of an event) need not be an exact copy of the original, rather a meaningful interpretation of the important points such that the original could be recognised by the representation. How much of the information and what information needs to be represented is extremely context dependant.

An everyday example of this is the representation of a landscape as a map. An aerial photo represents a huge amount of the information in the landscape, however, is too detailed and complex to easily interpret for the purpose of planning a journey. On the other hand, a line sketch containing the key features - stream/road routes and locations of buildings - is an abstract representation of the landscape, containing less of the information but is more useful to the journey-maker. In this way, refining the unstructured information contained within the text descriptions into a set of attributes reduces the complexity of the representation and allows analysis and interpretation; however, it is important to ensure the information-loss does not occur to such an extent that it becomes impossible to recognise the situation it describes.

'Frequent' attributes

Initially, consider those attributes which occur in >1% and fully describe 81% of annotated event descriptions. These are listed in Table 5.1 in Chapter 5. The high proportion of failure event descriptions completely described by this finite set of 'frequent' attributes corroborates the underlying premise of attribute-based analysis: a succinct set of core attributes can depict work site situations. While these 'frequent' attributes can be said to produce valid representations of the majority of situations, it must be considered whether their granularity is sufficient to prevent excess information-loss.

Discussed here are two points: the detail of the attribute classes themselves, and the applicability of these attributes to new data.

The detail level of the attributes identified during the annotation exercise was based on previous research at the University of Colorado, such as Desvignes (2014). These attributes provide specific object types, e.g. nail, but do not give additional detail such as size, composite material. In the majority of text descriptions, the 'type' level of detail was available in the text description while the added level of detail (size, material etc) was not. Therefore, this level of detail is appropriate for attributes representing the text description. Additionally, for safety events, this gave a good level of detail for further analysis, as seen in Chapter 6 where further analysis methods produce meaningful patterns and correlations.

However, aggregation of similar attributes at the end of the annotation exercise has created some attributes which appear to be more a category than an individual attribute. For example, attributes such as 'adverse weather' or 'machinery' could refer to several specific types which they contain, respectively 'storm, extreme cold, wind etc' or 'excavator, forklift etc'. Returning to the map analogy, this could be compared to including roads on the map but not the type of road - motorway, A-road or country lane.

These distinctions make a difference in the analysis. For example, by grouping large machinery together, further analysis is now unable to identify whether there is a specific risk, or attribute cluster, associated with one type of plant. However, without aggregation, these attributes would have been extremely infrequent and it would be unfeasible to include them into the analysis altogether, due to the inability of automatic prediction (see next discussion point 'infrequent attributes' for more detail).

Additionally, the description data itself is not detailed enough to consistently drill down to the sub-attributes level, with generalisations made by the reporter such as 'plant moving across walkway'. In this case, if attributes were included at the level of plant type, missing data would become a problem where an attribute which is present is not included in the attribute set.

There is, therefore, an important trade-off in the granularity of the unique attribute classes for representation and the detail required for the type of analysis desired. For an initial step into attribute-based analysis, the level of attribute granularity gained through the annotation exercise in this research is appropriate for representation of the event descriptions. This is because of the risk of missing/incomplete data at a finer granularity. However, future work could consider the use of an attribute taxonomy, where more detailed sub-attributes become features of the attribute. This is discussed in Section 8.2: Future research and Limitations.

On a tangential point, even if an appropriate structured representation of the text description is accomplished, not all the required information about the failure event is consistently contained within the text description. There is, therefore, a requirement to consider what other data and information is needed to represent the failure event - as opposed to simply representing the text description. The requirement to include other data in analysis is discussed in sub-section 7.1.4.

The next consideration is whether this set of 'frequent' attributes is also representative of data collected on other projects. In any data task, results should not normally be extrapolated outside the range of observation. The data used in the annotation exercise to create the training set was based on 28 infrastructure projects in 10 sectors. The wide variety of sectors and projects demonstrates a level of confidence in the applicability to general infrastructure projects.

There are certain caveats to this applicability:

- Specific sectors may have more emphasis on certain attributes. Such that, if the same number of descriptions were annotated from ONLY that sector, there would be attributes which would be 'frequent' which are currently in the infrequent category. To confirm this hypothesis, a huge investment in annotation would be required. This is most likely infeasible for every industry.
- These attributes are captured at a certain point in time. It is (hopefully) inevitable that new methods and materials become available and popular in construction. At which point, the set of 80 'frequent' attributes here will no longer be representative of the majority of work site situations. The next section, Section 7.2: Application of AI for learning from text-based failure data, discusses different ML and text analysis methods to ensure that the attribute list stays up-to-date.
- These data are all UK-based projects. Differences in 'frequent' attributes may occur in different territories and locations. For example, despite initiating the attribute annotation from the attribute list found in Desvignes (2014), a slightly different attribute set was identified in this research. The differences are extremely subtle, coming down to a couple of attributes, suggesting that that differences in core attributes between areas may be slight.

The final consideration to determine whether the results here can also provide some insight into development of this method for the analysis of the other two failures which have an 'onsite-point-of-failure': environment and quality. The list of 80 attributes here were identified given that a safety event occurred. Therefore, a point of contention here is whether the set of attributes found to represent the majority of work site situations which resulted in a safety event are the same as those to represent the majority of all work site situations.

It must be stated that, without further analysis using quality and environment data, it is not possible to conclude whether or not the existing set of 80 attributes is representative. However, by examining the context, it is possible to draw conclusions about the likelihood of their applicability. The data this research used to discover these attributes contained safety incidents and observations of unsafe situations. The inclusion of safety observations arguably increases the suitability of these attributes in comparison to those which resulted in a safety event, as it includes a more general representation of site environments. Additionally, to a certain extent, environment, quality and safety errors present are interconnected as, for example, incorrect quality or a spill can create an unsafe environment. Therefore, although the likelihood that this list is exactly the same is low; it is likely that these list sets are similar. However, returning to consideration of the attribute detail level required to represent these failure events sufficiently, quality and environment failures have different requirements to safety.

For environment failure events, there are certain attributes which would require greater granularity, such as 'hazardous substance' or 'vegetation'. This could be incorporated as a lower level of taxonomy or as a separate stage of analysis as, in this case, there already exist text processing methods to identify specific chemical names - mainly from the medical field - and therefore, this could be added as a separate stage in the data analysis.

The discussion of a taxonomy of attributes becomes more relevant for quality events (NCRs). For these events, an extra level of information, containing further information about the materials, would be needed to represent the situation to a reasonable extent. For example, it is not enough to identify a 'steel section' but the steel grade, dimensions and purpose (beam vs column). This higher granularity information should follow that contained in BIM models. By including the information in a taxonomy, starting from the general 80 attributes and drilling down to specific material grades and dimensions, more lines of inquiry are unlocked. The high level, low granularity set of attributes can be used for overall trends. Meanwhile, higher granularity attributes would allow engineering staff to drill into specifics. For example, the high level analysis can find a rise in the trend of NCR related to bolt fixing and the deeper dive can find whether this trend is related to a specific type of bolt, or a method/activity.

Infrequent attributes

Another aspect to consider is how to deal with infrequent attributes. Unusual attributes, like unusual map features, make a situation immediately recognisable. This added granularity of attributes could increase the depth of the representation of this method.

However, during the iterative annotation exercise, it was found that identification of infrequent attributes is more subjective than the identification of key attributes. Additionally, this research found that using supervised ML to automatically extract these extremely infrequent attributes from the text descriptions is untenable due to the low positive cases. For example, for the 113 attributes which were only identified once in the annotation exercise, it is impossible for any machine learning algorithm to discern any pattern which would allow classification. This task is the equivalent of trying to fit a curve to a single point.

There are three options to deal with these infrequent attributes proposed here:

1. Include them as is into the attribute set;
2. Ignore them from the attribute set; or
3. Group them under a 'contains an unusual attribute' attribute.

If it were possible, the first option creates a 'higher resolution' representation of the event. By including many extremely infrequent attributes, the representation may be immediately recognisable of the situation. Also, as previously mentioned, there are subjectivity issues with manually identifying these attributes and later automatically extracting them using ML methods.

Additionally, in considering the purpose of attribute-based analysis - that is to draw upon the collective knowledge of multiple events - the inclusion of infrequent attributes is irrelevant to knowledge discovery. Consider the analysis methods as presented in Chapter 6. Here, significant relationships and risks are being observed. Therefore, the inclusion of an attribute occurring extremely infrequently in the data is not statistically significant, rather producing a large amount of noise for the analysis method.

Therefore, inclusion of these less frequent attributes as unique attributes is currently unsuitable, due to the inconsistency in identification during annotation and issues in automatic extraction for supervised ML. Future inclusion of these attributes, should new methods allow their automatic identification, should consider the trade-off between representation detail and significance for analysis.

The second option, which was adopted in the analysis presented here, results in incomplete representation of approximately one fifth of descriptions. However, consider, does the absence of these infrequent attributes make the situation unrecognisable? Not in all cases. For example, take the following example from the data:

> *Checking for battery for PDA in security box by entrance in main building and slipped on banana skin.* [1]

In this case, inclusion of 'banana skin' as an attribute in the representation would make the situation immediately recognisable, if for the comedic value alone. However, without its inclusion, the attributes 'object on floor' and 'walking/moving around' still capture the key information about this safety event. Additionally, the collective analysis benefits from the inclusion of this more generalised attribute and is unaffected by the exclusion of the singular observance of 'banana skin' as infrequent attributes do not present statistically significant results to the whole.

The final option is proposed as it would allow management to assess the relative risk of undertaking an unusual task or using an unusual tool/material/object compared to those which are frequent. This may be useful as it could raise a flag for extra care or additional checks to take place prior to an unusual activity. However, when analysed quantitatively, any risk analysis would conflate unusual attributes which may be higher risk with those which are lower risk. The results of such an analysis would therefore be meaningless. Also, in the experience of this research, identification of 'an unusual attribute' using text classification ML methods is extremely unlikely to have good performance using the presented text classification method.

To conclude, inclusion of infrequent attributes would not add significant depth to representation or benefits to the attribute-based analysis as presented in this research. It is possible that, as the volume of data grow, that a greater level of granularity may be appropriate. At that point, inclusion of more infrequent attributes should be considered, perhaps in parallel with taxonomy hierarchies, such that infrequent attributes can be aggregated or dis-aggregated as appropriate.

[1] This is a genuine incident report from the data used for this research. Whether it is a genuine incident, however, remains to be seen!

7.1.3 How useful are the attributes refined through this research?

Given that the 'frequent' attributes extracted from text descriptions of failure events create a representative set of structured data but miss the nuances of the situation, is the analysis unlocked by using this set of 80 attributes useful to the industry? This discussion returns to validate the theorised advantages of attribute-based analysis in light of the results of this research. These claimed advantages are: increased granularity of quantitative (risk) analysis, explainability of result and anonymisation of data.

Increase granularity of analysis

Example methods in Chapter 6 illustrate several of many ways in which attribute-based analysis can investigate failure events. The methods proposed here apply equally to quality and environment failure criteria, given that representative attributes associated with these failure events could be extracted.

The demonstration of quantitative risk analysis in Section 6.2 validates the claim that attribute-based analysis facilitates a greater granularity of risk analysis. Meanwhile, the demonstrations of further analysis in Sections 6.3: Method 2: Use of predict incident outcomes and 6.4: Method 3: Use of attributes for network analysis show other quantitative analysis made possible through attribute-based data.

A key consideration, when employing these quantitative analysis methods, is to identify significant levels of attribute granularity for knowledge discovery. As previously discussed (see 'banana skin' example), inclusion of highly granular but infrequent attributes would result in an inability to identify meaningful patterns or relationships in the data. For example, if all attributes were extremely specific but only occurred 0.1% of the data, there would be few occurrences and fewer co-occurrences to form a network as in Section 6.4. At the other end of the scale, highly aggregated attributes - e.g. 'a material' - would produce equally meaningless results.

Identifying the level of granularity required for these analysis methods to be useful is, therefore, an iterative process. Previous discussion explored the level of granularity required for the attributes to be representative, and demonstrated that comparison of granularity levels used in previous research is a useful indicator. The same approach is explored here.

A previous case study by Hallowell and Gambatese (2009a) explored the risk profile of different aspects of a single 'task', in this case concrete formwork construction. Hallowell and Gambatese (2009a) used attribute-based analysis to show that different aspects of the task had significantly different risks. In achieving a differentiation of risk value between attribute risks, it could be said that this analysis is useful, as it allows construction professionals to allocate resources and focus mitigation techniques. In this way, the granularity of attributes observed in this case study is validated.

In a similar manner, the methods presented in Chapter 6 demonstrate meaningful patterns within their results which indicates that the granularity of the attributes used for these analyses were appropriate. The attributes used for Hallowell and Gambatese (2009a) were from the same set as those used to initiate the annotation exercise in Chapter 5. Therefore, unsurprisingly, the attribute granularity for both are extremely similar.

Explainability

The ease of explaining ***why*** a model or analysis method has come up with a certain answer is essential to the construction community (see Chapter 4) and the consideration of different AI methods in light of this is returned to in the nest section, Section 7.2: Application of AI for learning from text-based failure data. However, before addressing explainable methods for automatically extracting these attributes, it should also be noted that the choice of attribute-based analysis is itself a methodological choice to increase the explainability and meaningfulness of analysis. By extracting attributes as a way-point between the unstructured text and any down-stream task/analysis, interrogative techniques to explain the results of the analysis are more meaningful.

An example from this research is the use of event descriptions to predict the outcome of the event, presented in Section 6.2. In literature, such as Zhong et al., 2020, where the text description is used to directly predict the outcome category, it is harder to explain the result directly to the user. While some effort is made to post-prediction to identify the most significant predictors in the form of unstructured phrases and words, the lack of structure and consistency does not allow these predictors to be used directly to further trend analysis. Additionally, 'predictive' elements in the text often refer to the outcome itself, not valid precursor attributes.

However, with attribute-based analysis, it is possible to identify 'significant' predictors which are explainable and meaningful to construction professionals, as demonstrated in 6.1 to 6.3 in Sub-Section 5.5.2 for SVM prediction. Therefore, the experience of this research endorses the claim that attribute-based analysis increases the explainability of analysis using text-based data.

Anonymisation of data

In employing attribute-based analysis, based on a fixed list of attributes as in the text classification task presented here, complete anonymisation of the data is achieved. It is important to note that this is not true for other methods of attribute extraction, as discussed next in Section 7.2.

Anonymisation of failure event data is important because of the highlighted psychological and commercial issues with publicising failure information. If a consistent, anonymised representation of the data could be formed, this would facilitate data sharing across the industry, allowing industry-wide trends to be observed.

For this reason, while a fixed list of attributes has draw-backs (for example, subjectivity of inclusion and 'known unknowns'), a set of industry standard pre-defined attribute classes would generate the most impact to the industry as a whole.

7.1.4 What implications does this research have for collecting data?

The results and analysis presented throughout this research have revealed several key implications collection of data about failure events. The main theme for this discussion is data epistemology.

Return, for a moment, to the discussion on epistemology presented in Chapter 3. Each data type collected has firmly embedded within it an epistemological assumption, which outlines how you are able to discover or inquire about reality. Certain types of data encourage, perhaps falsely, that the information they contain be interpreted in a certain manner and represents a certain type of 'fact'. For example, when presented with numerical data, society is quick to assume these data as a strongly positivist measure - a measure of a single truth. While all manner of caveats and cautions may be associated, in reality, numbers "seem to be immune from theory or interpretation" (Poovey, 1998). This is also true to some extent for other types of structured data, including categorical data.

It is therefore essential in collecting or creating structured data, which will be used to implement change in industry, that these data be as closely aligned to a positivism stance as possible. When discord occurs between the assumed ontology and reality, the door is left open to misinterpretation. At best, this could lead to wasted effort and resource. At worst, it could lead to catastrophic failures.

At its core, this research aimed to draw on the collective knowledge contained within descriptions of failure events. A key task for this is to convert the unstructured text data from failure events into a structured data form which could be used in digital analysis. This transforms narrative data - which encourages sense-making and understanding of the individual event - into structured data - encouraging identification of correlations and quantitative analysis of the collective. Three points are raised by this: (1) should the attribute data be collected directly rather than extracting this from text descriptions; (2) how should the attribute categories be fashioned so that they do not invite misinterpretation; and (3) what other data/information are required to complement attribute data.

Should attribute data be collected directly?

A key finding from the semi-structured interviews in Chapter 4 was that additional manual data collection was not suitable due to the current high cognitive load and time pressures on site personnel. Therefore, collection of attribute data may require the exclusion of another type of data. Perhaps the most obvious data to be replaced would be the text description itself. However, while the text description is not useful in its current form for analysis of the collective, it is key to facilitate sense-making into the individual event. This is especially important for those incidents which result in an injury.

Also, direct collection of attribute data poses another set of difficulties. The number of attribute categories would mean that any multiple choice selection is unwieldy as users try to find the correct one. Therefore, collecting categorical attribute data directly is not suitable.

How should attribute categories be fashioned so that they do not invite misinterpretation?

To ensure the attribute data set does not invite misinterpretation, attributes should be physical, objective and easily identifiable. This principle is needed because attributes claim the existence of something in a set of structured, categorical data. These attributes should be as close to objective as possible - i.e. they are either there or not - with little/no scope for opinion. This was a core principle used by our annotators during the annotation process to identify attributes in the following categories: objects, actions, site environment and personnel descriptors. These categories were identified from Tixier et al. (2016b). Despite this, upon further reflection, some attributes were identified which did not abide by these principles. These were all from the 'person descriptors' category and included:

- Improper body position
- Improper procedure
- Improper security of materials
- Negative human influences
- No / Improper PPE

These categories, while presenting themselves as objective, contain within them a level of subjectivity where the person reporting or recording the event has placed judgement upon whether something was 'improper'. The inclusion/exclusion of the 'No / Improper PPE' attribute in the representation was the hardest to decide upon. However, I decide that this attribute is more accurately collected as part of the immediate cause than a precursor attribute. For example, consider the following fictional anecdote: "a site worker has parked their vehicle in the site car park and is walking to the changing rooms when they trip on some uneven ground. At this time, they are not wearing any PPE." In this case, the attribute 'no / improper PPE' would be both true and not true. We want to capture the information of this type when it contributes to the incident - i.e. is part of the immediate cause.

What other data/information are required to complement attribute data?

Chapter 6 demonstrated how attribute data can facilitate understanding about failure events. However, there are other data and information which are (or could be) used to realise this understanding. Discussion here addresses outcome data, currently used in literature and industry. In the conclusions, other 'passive' data and future data sources which could unlock further levels of understanding are proposed.

Outcome data

For the safety data considered in this research, outcome data is collected in the form of incident categories i.e. 'types', body part(s) injured, injury sustained, and lost hours/days. These categorical data types are generally common across the industry. Further outcome data, which are unstructured text data, include: immediate action taken, root cause details, investigation summary and suggested improvements.

When considering collection of these data sets, it is important to identify why they are being collected and what further decisions or actions they aim to prompt. Referring back to the implied etymologies of certain data types, structured data collected about failure events is generally suitable for collective analysis and trend identification. Meanwhile, text data is suited for sense-making about the individual event.

Incident categories are used currently within industry to create dashboard overviews of safety events reported, as well as being used in research to investigate common risk factors e.g. Sun et al. (2020). This is an important source of outcome data and it is essential that it is both standardised across the industry and consistent in its collection.

For those reports of events resulting in an injury, the sub-categories are fed by industry standards - led by HSE in the UK - and are suitable, immediate cause classifications. In full, these were:

- Contact with electricity
- Contact with moving machinery
- Driving at work

- Exposure to fire
- Exposure to or Contact with harmful substances
- Fall from height
- Falling dust / debris into eye
- Handling, Lifting or Carrying
- Hit/Struck by moving or falling object
- Hit/Struck by something fixed or stationary
- Physically Assaulted / Injured by person or animal / insect
- Slip, Trip or Fall on same level
- Struck by moving vehicle

These sub-categories indicate immediate causes, where an 'immediate cause' is the event which directly resulted in an injury. In these cases, it should not be subjective whether this category is correct or not. The event either happened, or it did not. The same is generally true for body part injured and injury.

However, this objectivity and clarity does not carry to near-miss/observation reports which contain a confusing mix of immediate causes, speculative root causes (e.g. 'procedural deficiencies and shortfalls') and attributes (e.g. 'Plant and Vehicle Movement'). Also, subjectivity in identifying these categories was evident in the number of times the annotators disagreed with the category given in the original data. To some extent, the existence of these observation reports are subjective as site personnel have to recognise and judge something to be 'unsafe'. However, this should lead to sub-categories which are as objective as possible, preserving the integrity of this structured data type.

Ideally, the same categories used for events resulting in an injury would be used for near-miss/observations to allow direct comparison, however, it is clear that these are unsuitable as they imply, in most cases, that an injury has occurred. Perhaps, a suitable set of categories would follow the principles behind the descriptors of 'Execution of Activity' from Smith, Sherratt, and Oswald (2017)'s analysis of unsafety on construction, as seen in Figure 7.1. These should be objective descriptors of the execution of the activity, and should not contain the 'decision or trigger' categories which precede them. Identification of these 'triggers', in the moment, by personnel on site with personal connections to the event, do not lend themselves to be objective. Their identification is more suited to post-investigation. On closer observation of the categories included in Smith, Sherratt, and Oswald (2017)'s analysis, some could be considered slightly subjective - such as 'use of inappropriate equipment', therefore, further research is required to refine these categories and assess their generalisability before implementation as categories for data collection.

While the data sets for the outcome of environmental and quality failures differ, the principles of data collection outlined here hold true. The data types collected should be consistent with their epistemological implications and desired use.

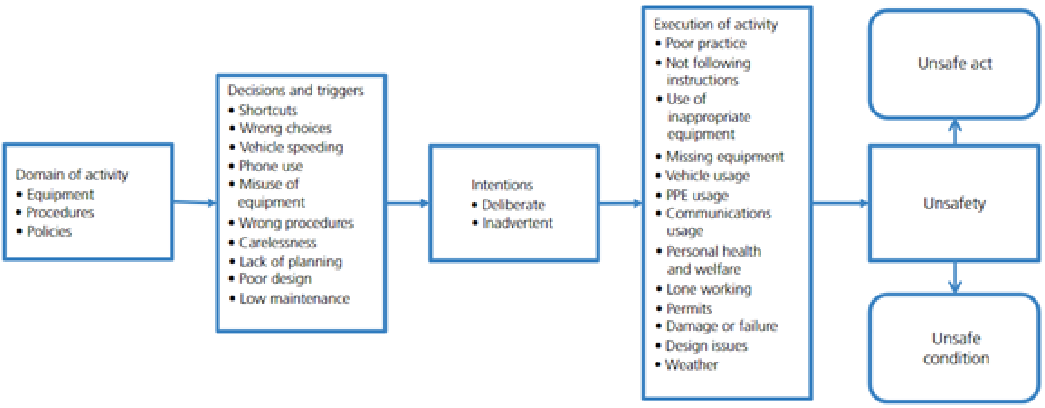


Figure 7.1: The development of unsafety. Image from Smith, Sherratt, and Oswald (2017).

7.1.5 Summary

The results gained through this research support attribute-based analysis for the investigation for events in the construction industry. Attribute-based representations of the text descriptions of failure events unlocks methods to analyse the collective, rather than sense-making individual events. However, this methodology should be supported by revising data collection including revising and standardising event categories.

Key findings include:

1. Frequent attributes create a valid representation of failure events to be used in further analysis. For the research presented here, using UK safety data, frequent attributes were defined as those occurring in >1% of descriptions.
2. Infrequent attributes were found to be identified more subjectively by the annotators and would not provide benefits to the analysis methods as they were not statistically significant.
3. Current granularity of attributes is useful for analysis given the level of data available. If more data become available, or quality and environment data is included, attribute taxonomies should be investigated.
4. Attribute data should not replace text descriptions of failure events. These two data types unlock different information: sense-making of individual events (text descriptions) and correlation/comparison (structured attribute data).
5. Event categorisation needs revising in light of epistemological considerations and standardisation.

7.2 How to choose a NLP & ML methodology

Having established that attribute-based analysis is relevant and required to exploit the learning potential of site-based failure events, the next step is to establish how these attributes should be generated. The main driver which this discussion circles back to is trust.

The need to build trust in these methods has been well-established in this research. However, when considering different approaches to AI in industry, two aspects of this are particularly important: trust in how the results are being calculated and trust in the results being accurate. This can be distilled to explainability and performance of the model respectively.

An interesting take on these issues is in a Forbes article *4 Unique Challenges Of Industrial Artificial Intelligence* which notes that "Technicians who have been in the field for 45 years will not trust machines that cannot explain their predictions." (Yao, 2017). Also, in addition to expecting the ability to explain AI predictions, industry expects and requires a much higher fidelity of model than consumers. In the same article, Harel Kodesh, while working as CTO of GE Software, is quoted to state that "In consumer predictions, there's low value to false negatives and to false positives. You'll forget that Amazon recommended you a crappy book". This mean that companies like Amazon and Google can deploy consumer-facing models with far lower performance values, then use these to collect data to improve their predictions, without customers rejecting their methods. In contrast, an incorrect prediction by an industrial AI can have devastating consequences for the trust in the model and its future use.

As with many important criteria, these two - explainability and performance - are often in a trade-off. As the algorithms get more complex and nuanced, a higher performance can often be achieved. For example, recently so-called deep learning methods, such as neural networks, have been outperforming most other algorithms. However, these neural networks are difficult to understand and even harder to pin down 'why' a model has made a specific decision. Hence, they are often referred to as 'black box' methods.

For the methodology presented in this research, there are two steps which involve AI: (1) extraction of the attributes and (2) application of the attributes to knowledge discovery - for example, Chapter 6 presented a method of outcome prediction of the incident sub-category using the event attributes.

7.2.1 Extraction of attributes

Different AI methodologies for extracting attributes from raw text were discussed in-depth in Section 5.3.2 in order to inform the choice of methodology for this research. Those discussed were: rule-based NLP, keyword expansion, predictive region identification, identification akin to named-entity recognition (NER) and whole text classification. Text classification, as selected for this research, then had two further decision points: how to represent the text as a numerical vector and which classification algorithms to investigate. This discussion re-visits these decision points in light of the experience of application of those decisions.

Is text classification the correct methodology to extract attributes?

Supervised text classification has several advantages over other Natural Language Processing (NLP) methods to identify attributes from text data. Section 5.3.2 highlighted these as:

1. Re-trainable - Unlike rule-based NLP, text classification methods can be re-trained if more/new labelled data become available. This allows the inclusion of different attributes and updated phraseology/grammar.
2. Accounts for semantic context - Unlike keyword expansion, which is essentially a 'smart search', text classification considers the entire text and can exploit 'clues', such as linking

meaningful phrases to attributes.

3. Applicable across failure types - In creating a structured set of representative attributes for construction site activities, these attributes can be applied across other failure types to create compatible data sets.
4. Only includes event precursors - A drawback of predictive region identification is that the algorithm often 'cheats' by using description of the outcome in the text to predict the event category. By requiring a structured set of attributes, this is eliminated.

In the experience of this research, text classification has proven itself to be a suitable methodological choice for this task. Chapter 5 demonstrated that application of this method can achieve agreements up to 75% of the level achieved by human annotators using only coarsely tuned hyper-parameters. Meanwhile, application of this methodology to the entire data set in Chapter 6 created a representative set of attribute vectors (as assessed using using two-tail paired sample t-test) which could be used in further analysis to gain valuable and meaningful results for construction professionals.

This methodology is characterised by two tasks: converting the text data into a numerical representation and applying a classification algorithm to predict the attributes using that numerical representation. Both these tasks involve choices which impact the explainability and performance of the method.

An interesting future avenue could consider text classification as part of a pipeline with application of an NER-type task. This is considered further in discussion of 'How to decide which algorithms to explore?'

Is 'Bag-of-words' the best way to numerically represent these text data?

Research prior to this investigation and results published throughout have failed to realise significant advantages of employing more sophisticated text representation than vector representation, i.e. 'Bag-of-Words'. These long, sparse vectors, based on simply counting the words present, are also explainable which confirms their suitability here in light of the underlying principle of trust.

However, this representation model has probably been fully exploited and future advances in this space are likely to be slight. Limitations, such as word order loss, can be mitigated. For example, a recommendation in the application of 'Bag-of-Words' to construction text is to include frequent bigrams, as these are often essential in differentiating different types of construction activities, objects and worksites, e.g. "exclusion_zone". Nonetheless, word embedding models can encode far richer semantic details.

Recent advances in word embedding models, such as BERT language model (Devlin et al., 2018), have been shown to have large advantages in performance over the bag-of-words representation for many applications worldwide. Although using deep learning methods reduce the explainability of the text representation, future research should look to utilise the advantages to text analysis tasks these embedded vector representations bring.

How to decide which algorithms to explore?

For identification of attributes from text descriptions, human users can easily verify the result. For example, Table 7.1 shows three examples of the text and attributes predicted by the SVM Classifier. When presented side-by-side a human can verify whether the AI has got an attribute wrong or is missing any. This ability to common-sense check the resultant predictions reduces the requirement for the use of explainable AI for this step. At the same time, it increases the requirement for performance in order to build trust in the model.

If the decision point for deploying such methods is the development of trust in the system then

Table 7.1: Examples of attributes predicted from descriptions by the SVM algorithm

Water froze in a tube which as the frozen water expanded lifted the tube up from the spigot of the handrail to a set of stairs	Objects: Water, Guardrail, Stairs Worksite environment: Adverse weather
Tippers and bobcats operating without a banksman.	Activity: Driving Object: Machinery
While moving a toolbox through doorway in Southern tunnel, IP trapped little finger on right hand between site box wheel and door frame. IP sustained a bruised and swollen tip of finger and slight bleeding from around the nail. IP received first aid treatment and continued work.	Activity: Lifting Object: Door

a sensible performance to achieve prior to deployment might be where the computer AI achieves agreement with the annotated training data equal or close to the agreement of human annotation. Currently, the text classification task presented here operates at 75% of this threshold.

Another consideration for application is the prioritisation of recall vs precision. Human users are able to reject incorrect attributes far more simply than they can identify missing ones. This indicates that recall is more important than precision. However, returning to the development of trust, if incorrect attributes are consistently predicted, users will lose trust in the system. Therefore, it is suggested that both measures are equally important so F1 (the harmonic average) is the correct performance metric to maximise, as is used here.

Having established a lower requirement for explainability, this opens up the possibility of more complex ML pipelines and algorithms to increase performance.

A key feature of this methodology is the ‘plug and play’ nature of algorithm adoption. Classification algorithms are developing extremely quickly. To pick a methodology which “locks in” a particular type of of classification would be unwise and, in all probability, hinder future developments. Therefore, the flexibility to change and upgrade algorithmic choices is essential to this methodology.

At the same time, regardless of the future developments of the algorithmic choice, it should be appreciated that the construction industry does not currently possess high levels of machine learning expertise, therefore, the sensitivity of algorithms to uninformed human input and overfitting should be minimised to a reasonable extent. This does not mean that the methods should be ‘dumbed down’ rather that systems which employ such technologies need to be extremely carefully created.

This is especially true in the case of ‘AutoML’ classifiers. Google’s cloud-based ‘AutoML’, released in January 2018, aims to “allow firms with limited data science expertise to develop analytical pipelines capable of solving sophisticated business problems” (Abbasi, Kitchens, and Ahmad, 2019). Applicable to the construction industry, these algorithm collections, which automatically tune hyper-parameters and select the best performing algorithms, could unlock huge potential in these, and other, data across the industry. However, there are risks in encouraging this type of data analysis by under-trained personnel, most of which revolve around how to deal with bias: bias in the data collection, selection of training data and interpretation of results. This research found significant indications of bias in the data, which would, perhaps, have been overlooked using automated pipelines. Therefore, this research concurs with Singhal (2019)’s recommendation that “more research must be done into the theory behind machine ethics and its implementation and better understanding how ML models make decisions/inferences and the impact thereof".

A method to increase the performance of the text classification could be to reduce the 'noise' in the text data. An initial step to identify the text referring to the attribute categories - actions, objects and site environment - could narrow the tokens included to the text classification task. This would require including an 'NER-like' task before text classification in the processing pipeline. However, this would eliminate contextual clues to the attribute - for example, the tokens 'at the top of' would not be identified in relation to an object but could be related to ladder or stairs.

Another option, instead of text classification and if the token set were succinct enough, is clustering the resultant NER tokens then automatically labelling the clusters. Automatic labelling would not guarantee anonymisation as a certain project name may be incorrectly deemed an attribute, however, would allow new or unknown attributes to be automatically identified and included. This would also eliminate the requirement for more labelled data.

To address the lack of anonymisation, it may be possible to label these clusters using a 'kNN' task with previously identified attributes. In this manner, sectors would be be able to manually add attributes to their 'frequent/include' list, without risking the anonymity of the attribute representation.

7.2.2 Knowledge discovery using attributes

To develop trust in attribute prediction, this research concluded that the explainability of the method can be sacrificed to improve performance. However, the same does not hold true for knowledge discovery and further analysis. These tasks are those which can instigate change and therefore it must be possible for construction professionals to interrogate *why* an answer is given.

This discussion does not go on to recommend specific knowledge discovery methods, nor is it appropriate to do so. There will be different models appropriate for different knowledge discovery tasks. What is considered here are core principles, revealed through this research, which should guide future development of specific knowledge discovery systems. These principles echo those already explored: trust, transparency, usefulness and appropriateness.

Appropriate knowledge discovery methods, in the experience of this research, take a human-centred approach to design and development, acknowledging the biases in failure data and selecting methods which are suitable for the implied ontology of the data type and analysis question. They should also focus on developing trusted and transparent processes which complement the socialisation learning processes already in-place in construction.

AI methods selected for knowledge discovery should be useful. This research was built upon a philosophy of pragmatism. It aimed to create useful methods for learning from failure events. Section 7.1 established the usefulness of attributes to analysis of failure events, however, in considering the choice of methods for knowledge discovery tasks, this needs to be taken one step further and the user needs to examine what method will create the most *useful* analysis for a particular question. For example, the methods presented in Chapter 6 are useful for different tasks. Quantified risk (Section 6.2) has possible application in planning future activities and bench-marking for performance, while complex network analysis (Section 6.4) could be far more useful for sense-making of correlations and patterns. Therefore, construction professionals need to formulate their data analysis requirements, i.e. ask "*why am I performing this task? What question am I aiming to answer?*", prior to designing the knowledge discovery task and selecting methods.

7.2.3 Summary

This discussion has confirmed that application of text classification to identify and extract attributes from text failure data, as experienced in this research, is suitable when considering the context (Chapter 4). Exploration of this AI methodology has been invaluable when identifying some key findings for selecting AI methods for attribute-based analysis for learning from failure in construction:

1. Explainability can be sacrificed to improve performance for attribute prediction as human intelligence can clearly see the link between input (text paragraph) and attribute (key material).
2. Flexibility for the methodology to update and retrain the ML algorithm selected is key due to the pace of development in ML discipline.
3. F1 is the correct model performance metric to maximise for algorithm selection as both recall and precision are important to this application. An F1 of 66% matches the agreement between human annotators.
4. Future research should explore word embedded text representations and more complex pipelines for extraction involving text classification in conjunction with NER-like tasks.

However, it is also essential that deliberate, human-centred design is undertaken for creation and deployment of knowledge discovery methods using these attribute-based representations. Key principles for this are to select explainable methods which are both appropriate for the context of use and useful in answering the aim of the knowledge discovery task.

7.3 How do these findings apply to learning from failure in industry?

The previous two sections established the suitability of attribute-base analysis of failure events in the context of the construction industry, and set out principles to consider when identifying NLP-ML pipelines to extract these attributes from event descriptions. The final research question 'How is best to implement this type of learning into systematic processes for the construction industry?' also implies exploration of how attribute-based analysis and various methods of analysis, such as presented in Chapter 6, would develop organisational learning for the construction industry beyond the personal competency-based, lessons learnt alerts currently employed.

Returning to organisational learning and knowledge management theory in Chapter 2, it is essential to consider key theoretical concepts in light of the results of this research. These are structured as: (1) synchronisation of feed-forward and feedback knowledge transfer; and (2) consideration of removal of barriers to learning.

7.3.1 Synchronisation of feed-forward and feedback

Previous research found that synchronised knowledge transfer up and down the management hierarchy is key to maximising the effectiveness of organisational learning (Morland, Breslin, and Stevenson, 2019). Therefore, this discussion theorises how attribute based analysis can be implemented to facilitate synchronisation of knowledge transfer from failure events. Here, knowledge transfer up the management hierarchy is referred to as *feed-forward* learning, while knowledge transfer down the management hierarchy is called *feedback*.

Figure 7.2 illustrates these organisational learning processes as communication from collective sense-making activities at each level up and down the management hierarchy in UK construction. Their findings reflect the high reliance on human communication which was also highlighted in Chapter 4. Morland, Breslin, and Stevenson (2019) also conclude that a large weakness to this current model is the reliance on the timeliness and selection of this, predominately human-human, communication. They note that the different levels of business "rotated" through their learning cycle at different speeds, resulting in disjointed knowledge transfer, and that "*successful coupling required additional effort and precision timing*". At the same time, reliance on human intervention to communicate the outputs from these sense-making activities led to, at times, "*selective or embellished*" knowledge transfer.

Collective sense-making in Morland, Breslin, and Stevenson (2019)'s model is triggered by some form of observation or event, similar to the event which triggers single-loop learning (Figure 2.5 in Chapter 2). In centralising the communication of these observations across the different levels of organisation and automating the trigger for sense-making activity from these data, the learning cycles from these events could be synchronised. Attribute-based analysis in this research has been shown to create anonymised, structured representations of events. These representations can then be stored in a common data environment for analysis at each organisational level. Each level of the business is then able to undertake collective sense-making activities appropriate for their level of business. Sense-making in this context is understood to relate to the "analysis" node in the single-loop learning cycle.

Collective sense-making for feed-forward learning can be described as tasks which summarise and analyse data to better inform upper management or upstream/future tasks. The higher granularity analysis provided by attribute-based analysis, rather than reliance on "headline" event categories, allows upper management (such as project managers and executives) to gain more nuanced understanding of activities on site, and identify specific issues and risks early.

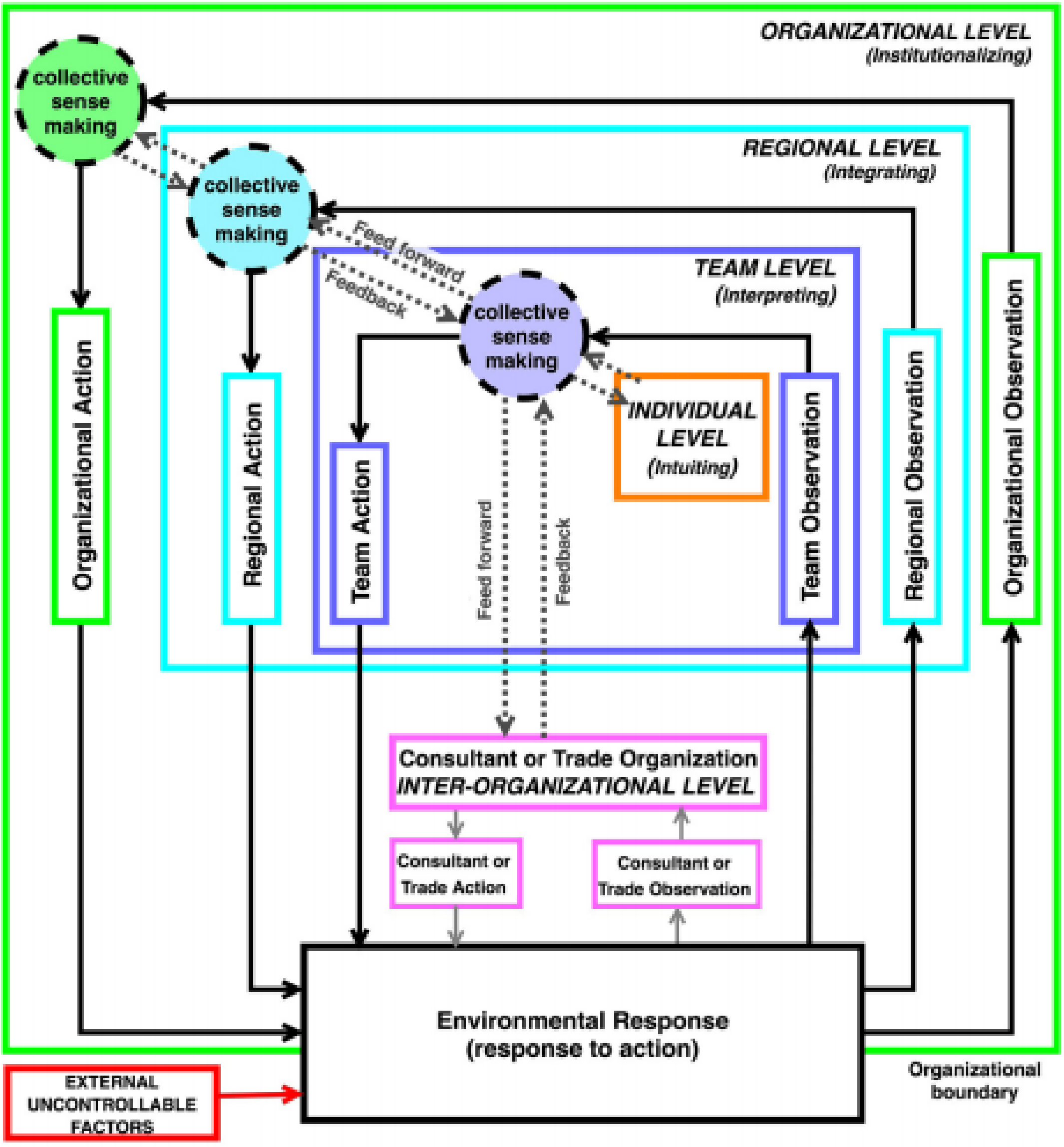


Figure 7.2: Multi-level learning model in construction organisation. Image from Morland, Breslin, and Stevenson (2019).

Feeding this data forward to infrastructure designers and clients can inform decisions about risks involved with different design and construction method options. Additionally, regulators could identify industry-wide trends and correlations, allowing government funds to more efficiently target 'problem' areas and create positive change.

Sense-making tasks, which feedback knowledge to site and project teams, should create curated and digestible information, relevant to the tasks and activities of the team. This is based on analysis from Chapter 4 where project teams recount the time-pressures and on Morland, Breslin, and Stevenson (2019)'s observation where "*individuals receiving the new knowledge were found to assess it for value against their level's experience prior to taking any action*". Attributes for new or upcoming tasks can be used to automate creation of risk profiles using the existing database. Additionally, recent failures resulting from similar activity profiles can be identified for inclusion into the planning process. In this way, the knowledge of the collective can be accessed and incorporated systematically as part of normal operations - not as a standalone task requiring time and effort to engage with - which was a limitation of the current learning cycle as identified in Chapter 4.

A possible drawback to this method relates to the increased independence of human-interaction and the knowledge transfer process. While reliance on human-human communication for knowledge has demonstrated many weaknesses, Morland, Breslin, and Stevenson (2019) cite selectivity and timeliness to name a couple, these interactions also ensure that relationships are built across the layers of business and help to communicate common values. Additionally, this method encourages sense-making to occur in parallel at different level of the business, rather than in series once the findings from one level are communicated forward/back. This could result in conflicting interpretations, which then percolate around the organisation. Efficient human-human communication, and communication processes, will be key to preventing this confusion. Therefore, implementation of data-based organisational learning processes should not replace human-human knowledge transfer - rather be used to enhance and complement these processes.

In this discussion, it is assumed that "best implementation" is a synonym for "most effective implementation". While this is clearly part of the evaluation for organisational learning processes, it does not assess the depth of learning or suitability to context. This returns to discussion on knowledge/data epistemology. Attribute-based data, and the analysis it entails, is suited to collective analysis of events to expose trends and quantitative correlations. This can be incorporated in systematic processes to assess risk and as input into continuous improvement processes, however, it does not lend itself to developing personal competence or in-depth analysis of individual events.

Therefore, attribute-based analysis for organisational learning could facilitate synchronisation of feed-forward and feedback learning for benchmarking, risk analysis and identification of complex attribute correlations. However, these methods cannot, in their current form, pick out underlying cultural issues on project or individual competency issues which are more suited to human-human sense-making activities.

7.3.2 Barriers to learning

This research is based on the hypothesis that attribute-based analysis can improve learning from failure processes in the construction industry. A way to substantiate this claim is to identify current barriers to learning which are removed or weakened by attribute-based research, as experienced through this research.

The barriers to learning from incidents which are discussed here were identified by Stemn et al. (2018) and fall under four categories:

- Learning inputs:
  - Non-detection and non-identification of reportable incidents
  - Under-reporting of detected incidents
  - Lack of focus on small precursor incidents
- Learning process:
  - Inadequate description of reported incidents
  - Superficial investigation and analyses of incidents
  - Poor selection, planning and implementation of corrective actions
  - Lack of effective learning from incidents (LFI) systems and sharing lessons
- Learning context:
  - Culture of blame, lack of trust and expected performance created by management
- Learning agents
  - Beliefs, experiences and competencies of actors of learning

Learning inputs

Perhaps the most immediate advantage of attribute-based analysis is the ability to collectively analyse smaller consequence incidents, including near miss and observation data, at a level of granularity which facilitates knowledge discovery. Currently, these data are limited in use to 'headline' dashboard activities, e.g. trends in numbers of observations, and there is a lack of motivation to expend resources investigating 'no consequence' events. However, in aggregating these events together, it is possible to identify trends of 'unsafety' and instigate actions from these small precursor events.

This greater granularity of data can also shine light on the possible biases contained within these data. For example, in Section 6.4 Chapter 6, the observation data demonstrated a higher proportion of 'unsafety' concerning vehicles/plant and vehicular movement than the proportion of incidents involving vehicles which resulted in an injury. Two possible hypotheses to explain this are: (1) there are proportionally more occasions of 'unsafety' involving vehicles than incidents which result in injury; or (2) 'unsafety' involving vehicles is less under-reported than 'unsafety' involving other attributes. Imagine that the second hypothesis is true. Observation of this bias may lead to the supposition that this renders this data useless for empirical analysis. However, in highlighting this observation, underlying site behaviours are uncovered, and new research avenues opened. For example, is 'unsafety' involving vehicles reported more because it is more visible? Because gate guards have time and opportunity to report more than workers on site? If the second is the case, should the construction project be making it easier for those on site to report 'unsafety' immediately? How does this affect safety in terms of using mobile devices? In this way, even 'flawed' data, can be used to gain valuable insights into the actions and behaviours on site.

This leads to the second barrier to discuss: under-reporting of detected incidents. It is possible that attribute-base analysis can be used to identify when incidents of a certain type may be under-reported, especially if attribute-base data is combined with programme and site diary data. However, the largest anticipated impact of implementing attribute-based analysis is a cultural shift in reporting by increasing the visibility and transparency of the collected data. Al-Aubaidy, Caldas, and Mulva (2019) found that under-reporting of construction incidents in the USA can be attributed to factors such as incentive systems, inadequate safety communication and unsuitable data management systems.

With the exception of financial/personal incentives for zero incidents, this boils down to a simple principle: If people are able to see the purpose and worth of a data collection process, they are more likely to participate in it. A personal anecdote of this originates from my time on site. Here, observation data collected the day before was incorporated into the 'Start of Shift' brief the next day, with actions addressing the 'unsafety' if appropriate. When the site staff could see that their observations were being heard, and something was being done about it, the number and quality of observations increased. This system relied on effective digital systems which allowed collected data to be accessed immediately by site teams, as well as being aggregated for upper management. It also relied on consistent communication of these safety data. As efficient as this method was for short-term individual events, longer term trends and re-occurrence was still an issue. By implementing attribute-based analysis in a similarly transparent way, with direct feedback to sites, it is envisioned that this will also build trust and the perception of 'usefulness' of reporting incidents.

The final barrier relating to the learning input, as identified by Stemn et al. (2018), is non-detection of incidents. While this is less immediately affected by this method, by identifying possible biases in the data, attribute-based analysis could be used to formulate new hypotheses into the non-detection of incidents.

Learning process

The construction industry exhibits a high reliance on human-human interactions for organisational learning and an emphasis on human competence which is not complemented by effective learning from incidents (LFI) systems. The results of Chapter 4 concurred that communication of these lessons is highly reliant on an 'alert' system, where safety incidents are the most frequently communicated, and require construction personnel to have time, opportunity and inclination to synthesise and implement lessons from these documents. The conclusion of this chapter determined that the industry could benefit greatly from formally facilitating socialisation activities to increase the effectiveness of tacit learning and sense-making activities. Meanwhile, there was a need to develop effective LFI processes which used explicit knowledge. Attribute-base learning would aim to complement sense-making of individual events and develop integrated, data-based systems for continuous improvement.

A barrier which this research also struggled with was 'inadequate description of reported incidents'. Identifying attributes depends upon sufficient description of the event. In one way, this may prove a significant limitation of this method. If attributes are not identifiable from the descriptions, attribute-based analysis will suffer from missing values and incomplete data. On the other hand, having now identified what is required from a description to make it useful - i.e. the objects being used, the actions being performed and description of the work environment - these can be developed into specific and succinct guidelines to improve the quality of event description.

The next two barriers to learning are discussed concurrently: 'superficial investigation and analyses of incidents', and 'poor selection, planning and implementation of corrective actions'. At the beginning of this research, an issue was highlighted with selecting corrective actions from the exceptional events - those which resulted in significant, often catastrophic, results. The conundrum is that smaller consequence events do not instigate in-depth investigations, therefore, do not develop considered corrective actions or lessons learnt. Meanwhile, learning developed from significant case-studies has a track record of exposing organisational failures at a high level, but does not identify actions to correct the multitude of small failures at the productive activity and defence barriers, i.e. the 'leaves' in Section 2.3.

Attribute-based analysis can bridge this gap. By allowing aggregation of similar events, cumulative consequences can be observed and justify more in-depth investigations. Meanwhile, consideration of multiple events during analysis of these failures can lead to a more holistic approach to identifying corrective actions.

Learning context & Learning Agents

A significant section of the research has ruminated on the context of learning in the construction industry. Of those factors identified by Stemn et al. (2018), constant themes were trust, blame and the experience of the actors of learning.

Trust and blame were identified as different sides of the same coin. Chapter 4 found that a reluctance to take ownership of failures and the prevalence of 'blame culture' was suppressing learning opportunities, innovation and collaboration. Meanwhile, increased trust - in processes and people - co-occurred with more complete learning cycles and effective communication. Implementation of attribute-based learning from failure targets sources of distrust, such as secretiveness and selective knowledge transfer, by developing a method which creates structured data which can be centralised and used at different levels of the business for appropriate tasks.

Therefore, the expected experience of those involved in the learning cycle - essentially everyone in the organisation - to increase trust and minimise blame culture, heavily influenced the methodological choices in this research. It should also influence the implementation of these findings into industry. The recommendation from this research is to approach the implementation from a socio-technical standpoint in order to deliberately incorporate the influences of the human into the system.

7.3.3 Summary

This discussion applied the findings of this research back into the realms of organisational learning from failure, where this research started. It concluded that attribute-based learning will facilitate effective systematic learning processes across levels of construction organisations by both creating opportunities for synchronised feed-forward/feedback learning and by removing or weakening current barriers to learning from failure.

Chapter 8

Conclusions

"A story has no beginning or end: arbitrarily one chooses that moment of experience from which to look back or from which to look ahead"

Graham Greene's 'The End of the Affair'

I could say that this section concludes this piece of research. However, like the sentiments behind this quote, I believe that this is just part of story which is on-going. Therefore, this section summarises and concludes the experience narrated so far. It then looks to the future, identifying applications and further research avenues leading from the outcomes of these investigations. I found this invigorating as I found myself getting excited about the future, rather than dwelling on the past.

8.1 Summary and Conclusion

8.1.1 Summary of research

This research evolved from a simple problem statement: the construction industry needs to learn better from its mistakes. To address this problem, four research questions were formed:

1. How does the construction industry currently learn from failure?
2. What recent AI and data science methods have been used in the construction industry, and what other methods exist?
3. Which Natural Language Processing (NLP) + Machine Learning (ML) model best facilitates knowledge discovery from text-based failure data?
4. How is best to implement this type of learning into systematic processes for the construction industry?

A pragmatic philosophy was adopted to guide decisions on methodology and method used to answer these research questions. This philosophy encourages the researcher to ask at each milestone 'what is the most useful way to carry out this task?'. By applying this pragmatic lens, different methodological stances were identified for different sections of this research. Adopting different methods in sequence in this manner is referred to as a multi-approach (Johnston, 2012). Specifically, constructionism principles directed an initial qualitative investigation in Chapter 4 which explored the current state of learning form failure' in construction, while a post-positivism stance framed the exploration of ML and NLP technologies for structuring the knowledge within text descriptions of failure events on site in Chapters 5 and 6. Finally, the discussion in Chapter 7 branched back towards constructionism, re-contextualising the previous results.

The literature review revealed the potential of unstructured failure data to delivering insights for the construction industry. However, a lack of underlying understanding about the concept of failure and learning from failure in construction currently hinders application and further exploration.

Several theories and concepts from existing literature were key to frame decisions and discussion. From the disciplines of organisational learning and knowledge management, the two which are central are Argyris (1977)'s process of 'Single vs Double Loop Learning' and Tuomi (1999)'s 'Hierarchy of Knowledge'. These core theories can help to structure and integrate exploration of other literature, such as barriers to learning (e.g. Stemn et al. (2018)), types of organisational knowledge and processes to transfer/convert this knowledge (i.e. learning), and socio-technical considerations (e.g. Gammel et al. (2019)).

It was found that there is a lack of agreement and foundation to the definition of failure in construction, especially when limiting the literature to a UK context. In particular, failure seemed to be often defined as the absence of success. Success literature, including topics such as identification of success criteria and factors and assessment methods, appear to be directly applied to failure. However, there was no evidence found to support the belief that the criteria and factors to achieve success are equally important to avoid failure. In fact, the psychological, cultural and human factors literature suggest that this assumption could be incorrect or at least require investigation.

Therefore, to address this literature gap, a qualitative piece of research, presented in Chapter 4, was undertaken. This qualitative research took the form of 19 semi-structured interviews with members of the UK construction industry, which were then thematically analysed with the aid of NVivo software. The investigation aimed to uncover details about the prioritisation of different failure criteria in UK construction and understand the existing learning processes. This understanding was essential to the success of this research.

Chapter 4 presented the qualitative research piece investigating failure and learning from failure in UK construction. Nine perceived failure 'modes' (i.e. criteria) were identified: Time, Money, Health & Safety, Public Perception, Stakeholder Management, Structural Collapse, Design, Environmental and Quality. Of those identified, the first 3 were highlighted as the key modes with one interviewee stating that "*the others all feed into these three*". For these failure criteria, analysis of the interview data showed different stages of maturity in the learning cycle. While safety showed mature single-loop systematic learning and some migration towards double-loop thinking, quality presented an undeveloped single-loop process. Time and money failures gave no indication of any systematic learning process; however, there was strong evidence of informal learning and discussion.

This investigation also revealed a high dependency on human competence and knowledge transfer via the passive occurrence of socialisation, i.e. human-to-human knowledge transfer through spontaneous or natural conversation (Nonaka, 1991). This reliance seemed firmly embedded into the culture of the construction industry. Therefore, further organisational learning processes should look to enhance or complement this socialisation - not replace it. For example, there are two ways that the construction industry can immediately look to improve tacit learning from failure which are support by the results of this investigation. These are: develop more formal socialisation processes and to rethink lessons learned processes. This is outlined further in Section 8.3.

Moreover, the analysis highlighted the current inability of the construction industry to fully exploit its written resource for learning from failure. Analysis of text descriptions of failure events (such as safety incidents, observations of 'unsafety', and quality NCRs) are restricted to manual techniques, which are resource intensive, inefficient and subjective. Therefore, while these documents can be used re-actively on project for remedial actions, in-depth analysis is rarely performed except in the case of large consequence events. Small consequence events, which collectively can have a large impact to a project, were unexploited. However, modern AI and natural language processing methods could facilitate systematic analysis of these reports.

This initial investigation also found that safety, environment and quality are the most documented failure criteria and that the data are collected and stored in a similar format. These data consist of reports on failure events which capture descriptive details of the individual events. They are generally captured via a form, either physically or digitally, and contain several different types of data, including structured categories and unstructured text descriptions of the event. Of the data collected from these events, the most pertinent information for learning is contained within the text fields documenting the event itself because of the high proportion of incomplete field entries and risk of subjectivity elsewhere in the forms.

Three further points, uncovered by the interview data, were key for development of the informatics methods outlined in further chapters. The first was the lack of incentive from management for individuals to participate in formal learning processes. While every interviewee espoused the need for learning from failure processes in construction, they also noted the lack of time and incentives to actually participate in such processes. Therefore, it was essential to ensure that learning processes integrate with existing processes, reducing the additional time requirement and mitigating against becoming a 'box-ticking' exercise. The next two point are key values - human-centred and explainable ML - which were key in selecting and developing an NLP + ML pipeline suitable for learning from failure.

Having established a requirement to exploit the knowledge trapped within the hidden descriptions of failure events, the research progressed to investigating which technology pipeline (ML + NLP) would be best facilitate knowledge discovery from these data (i.e. Research Question 3). Chapter 5 began by exploring different methods of structuring the unstructured text descriptions of failure events. Safety data were the most accessible, complete and reported by interviewees to be the most reliable data out of the three failure modes considered (safety, quality and environment). Therefore, safety data were used for this investigation.

Following work at the University of Colorado (e.g. Desvignes (2014) and Tixier et al. (2016b) and Tixier et al. (2016a)), attribute-based representations were selected as an appropriate method to structure the unstructured text descriptions. Attributes-based representation aims to develop a finite set of core features which objectively describe the work site and can be identified prior to work commencing (or an incident occurring). These attributes fell under the following three categories: objects, actions and site environment descriptors. The extraction of these attributes from the text descriptions was formed of two parts: development of the attributes through systematic labelling of the safety report data set, followed by application of NLP and ML to predict attributes in new safety event descriptions. This two-step approach is adapted from protocols developed and observed at the University of Colorado, Boulder (for example Tixier et al. (2016b)).

The data used in this research were gathered from a large UK infrastructure construction company. The primary data set consists of 14,266 safety incident and observation reports from a central Health, Safety and Environment database, recorded over a 9 year period. Of these, 3491 incident reports from 28 infrastructure projects in 10 sectors were labelled in an iterative process by four annotators - of which, 3244 reports were unique. Those which were duplicated are used to calculate inter-annotator agreement, which averaged at 49.2%. This value demonstrates a moderate agreement which is acceptable considering the number of categories (i.e. attributes) involved.

Manual annotation identified 250 unique attributes, of which 60 occurred in 1% or more of the data. Despite this high proportion of infrequent attributes, 81% of the descriptions were fully described using the set of 60 attributes. By which, this means 81% of text descriptions were not labelled with any of the less frequent attributes. Additionally, 113/250 of the unique attributes identified only occurred once in the annotation set.

The proportion of attributes (24%) which fully describe the majority of descriptions (81%) corroborates the “Pareto Prinicple” or 80:20 rule. (Sanders, 1987) notes that this principle, developed from the observations of quality pioneer Joseph Juran, allows for prioritisation of action as, for many phenomena, 20% of variables will account for 80% of the results.

However, manual annotation is time-consuming, and this task took over 300 solid work hours (not including breaks or moving from one description to another). Manual annotation is therefore not suitable for deployment. Automatic detection of these attributes is required to make this method viable in industry.

Different AI methodologies for extracting attributes from raw text were discussed in-depth in Section 5.3.2, Chapter 5 in order to inform the choice of methodology for this research. Those discussed were: rule-based NLP, keyword expansion, predictive region identification, identification akin to named-entity recognition (NER) and whole text classification. Text classification, as selected for this research, then had two further decision points: how to represent the text as a numerical vector and which classification algorithms to investigate.

‘Bag-of-Words’ or ‘vector space representations’ were used in this research to transform the unstructured text descriptions into numerical vectors. These representations are based on the numerical frequency of unique ‘tokens’ contained within the training vocabulary. ‘Tokens’ generally include words but also may also include punctuation or numbers. The resultant representation is a very long, sparse vector. This method was chosen as, despite the increased semantic information and complexity of word embedded representation (i.e. deep learning language models), research exploring NLP for construction text had not yet shown this translated into significantly improved task metrics.

Several classification algorithms were then investigated to predict the attribute classes from the numerical vector. The four algorithm types investigated were: Naïve Bayes, k-Nearest Neighbour, Decision Tree and SVM (Support Vector Machines). Two ensemble methods - gradient boosting for Decision Trees and bootstrap aggregating (Bagging) for SVM - were also applied to the final two algorithms respectively. F1 score, the harmonic average between recall and precision, was identified as the most appropriate performance metric.

Overall, SVM (F1 = 54.2%) and Gradient Boosting (F1 = 51.8%) achieved the best performance scores. There was no advantage to implementing SVM bagging. Due to it added complexity, being an ensemble algorithm, Gradient Boosting is less explainable, and it is also slightly lower performing than SVM. Therefore, SVM was used to predict the attributes for the whole data set (14,882 descriptions) for the analysis in Chapter 6. It should also be noted, however, that these classification algorithms are still only performing at 75% of human agreement scores. They would need to achieve 66% F1 to outperform human annotators, compared to SVM at F1= 54.2%.

Further findings extracted from the exploration of these results are:

1. The relative small volume of training data inhibits the potential of all methods investigated. More labelled training data would probable significantly increase the performance metrics.
2. Attributes with a higher number of positive examples, i.e. those which occur more frequently in the data, achieve higher classification performances.
3. Oversampling had a significant positive effect on approximately half of attributes predicted due to the increase in recall outweighing the decrease in precision. Oversampling for these imbalanced data should be employed in the future.
4. The type of attribute - object, action or environment - had no significant effect on the classifiers' performances.
5. Hyper-parameter values, as expected, have a significant effect on the model performance. For use in industry, optimising - but not over-fitting - these will be key.

Three knowledge discovery methods, more accurately described as 'information generating', were implemented in Chapter 6 to demonstrate the potential of attribute-based representations. These three methods were: risk quantification, incident outcome prediction and network analysis. These methods demonstrated only a small range of the approaches unlocked by transforming unstructured text descriptions of failure events into a structured set of fundamental attributes. The information gained via application of these methods included forecasting future risk as well as sense-making of current/past failure collectives. These results could generate tremendous positive impact for organisational learning in the construction industry.

By the culmination of Chapter 6, this research had established that, for small consequence failure events, there simply isn't the stimulus to invest time and money into investigative efforts to analyse the event information and draw out what change is required. However, there was a desire to more effectively use data currently collected about these events to facilitate systematic learning, for which the most pertinent data are unstructured text descriptions of the failure event. Moreover, ML + NLP techniques exist which can combined into novel pipelines and applied to structure these descriptions, using attribute-based representations, in a manner which unlocks further analysis methods for systematic knowledge discovery.

8.1.2 Conclusions

The conclusion of this research took the individual findings of each step and critically cross-examined them. Having established motive and novel means to exploit the potential learning from failure events, it was then necessary to explore whether the experience of attribute-based analysis in this research stood up to scrutiny in light of the organisational learning context, and form recommendations for industry.

This exploration was split into three sections: (1) exploration of whether this research supported application of attribute-based analysis for the construction industry; (2) extraction of methodological recommendations for applying AI for learning from text-based failure data; and (3) expanding these findings to developing systematic organisational learning processes for construction projects.

First, examination of the suitability of attribute-based analysis for construction failures concluded that:

1. Frequent attributes create a valid representation of failure events to be used in further analysis. For the research presented here, using UK safety data, frequent attributes were defined as those occurring in >1% of descriptions.
2. Infrequent attributes were found to be identified more subjectively by the annotators and would not provide benefits to the analysis methods as they were not statistically significant.
3. Current granularity of attributes is useful for analysis given the level of data available. If more data becomes available, or quality and environment data is included, attribute taxonomies should be investigated.
4. Attribute data should not replace text descriptions of failure events. These two data types unlock different information: sense-making of individual events (text descriptions) and correlation/comparison (structured attribute data).
5. Event categorisation needs revising in light of epistemological considerations and standardisation.

Second, discussion of AI methodology confirmed that application of text classification to identify and extract attributes from text failure data, as experienced in this research, is suitable when considering the context (Chapter 4). Exploration of this AI methodology has been invaluable when identifying some key findings for selecting AI methods for attribute-based analysis for learning from failure in construction. Specific recommendations include that:

1. Explainability can be sacrificed to improve performance for attribute prediction as human intelligence can clearly see the link between input (text paragraph) and attribute (key material).
2. Flexibility for the methodology to update and retrain the ML algorithm selected is key due to the pace of development in ML discipline.
3. F1 is the correct model performance metric to maximise for algorithm selection as both recall and precision are important to this application. An F1 of 66% matches the agreement between human annotators.
4. Future research should explore word embedded text representations and more complex pipelines for extraction involving text classification in conjunction with NER-like tasks.

It is also essential that deliberate, human-centred design is undertaken for creation and deployment of knowledge discovery methods using these attribute-based representations. Key principles for this are to select explainable methods which are both appropriate for the context of use and useful in answering the aim of the knowledge discovery task.

The final discussion applied the findings of this research back into the realms of organisational learning from failure, where this research started. It concluded that attribute-based learning will facilitate effective systematic learning processes across levels of construction organisations by both creating opportunities for synchronised feed-forward/feedback learning and by removing or weakening current barriers to learning from failure.

To conclude, this research used novel semi-structured interview data to discover information on the state-of-play for learning from failure in construction. This identified an inability to systematically learn from frequent but low consequence failures. To address this weakness, this research explored and developed an attribute-based methodology and NLP + ML pipeline to structure the text descriptions of these events so that they could be used in digital analysis. By refining the unstructured data in this way, the complexity of the representation was reduced which allows analysis and interpretation. Use-cases were presented to demonstrate how these new data can be exploited for knowledge discovery. Critical analysis of the previous results established the suitability of attribute-base analysis of failure events in the context of the construction industry, and set out principles to consider when identifying NLP-ML pipelines to extract these attributes from event descriptions. Finally, by considering how these methods would integrate into learning from failure processes in industry, this research demonstrated the potential benefits of attribute-based analysis and formed recommendations for adoption.

The final sections of this chapter concern opportunities for future research to address limitations and for immediate recommendation for industry.

8.2 Future research and Limitations

> *No research is ever quite complete. It is the glory of a good bit of work that it opens the way for something still better, and this repeatedly leads to its own eclipse.*

This quote, attributed in multiple online sources to a "Mervin Gordon", summarises the culmination of every research project. Unfortunately, no actual reference can be given to this quote as, on further investigation, it appears the only "Mervin Gordon" found by Google is a former New Zealand football player. However, the sentiment of the quote remains. No piece of research can explore every avenue open to it, and all pieces of research should open new avenues, even as they progress mankind's collective knowledge.

This research was no different. To explore those avenues which advanced the knowledge need for this research, constraints were put onto the research process to make best use of time, resources and expertise. Therefore, there are a number of limitations which should be addressed.

Additionally, throughout the journey, a number of new avenues were revealed which could not be explored (yet!). These are also included for consideration.

8.2.1 Limitations

Only safety data were examined

This project set out to examine learning from text-based failure data for construction, not just text-based safety data. Chapter 4 explored all different criteria for failure in the UK and their learning from failure processes, and Chapters 6 and 7 explored the general methods and applicability of attribute-based representations for learning. However, Chapter 5 adopted only safety data to illustrate and develop this method.

Safety data were used because safety was identified as the most significant form of failure in construction (in Chapter 4) and interviewees had the highest confidence in the completeness and uniformity of the data collected. Additionally, safety data were more accessible and had a greater volume of previous research to build upon.

However, this leads to several unknowns in term of developing the attribute set:

1. Do the attributes which describe most work sites/activities which result in a safety incident also describe those which result in a quality non-compliance?
2. Is the granularity of these attributes useful for investigation of other failure criteria, e.g. quality, environment?

Future research should address this limitation by developing an attribute set which also applies to other failure data. This could be done either via a manual labelling exercise (as is used in this research) or by development of semi-automated processes and taxonomies, explained in the next sub-section.

Data were from UK infrastructure projects

Another data limitation is the geographical and sector orientation of the data. While data used here represented a wide range of infrastructure sectors, extremely few buildings were included within these projects. Commercial property and house building projects were not represented, which represents a large proportion of construction in the UK. Additionally, the data were UK-based, and therefore could not be directly extrapolated to other contexts.

For application of these results outside of UK infrastructure, additional investigations would need to explore the contextual comparisons and possibly repeat sections of the data analysis to develop appropriate attribute sets.

Limited volume of labelled data

Manually labelling data is time-consuming. A compromise had to be struck between capturing enough, to develop the attribute list and for use in the ML processes, vs having the time to investigate the methods and results. There are two possible effects this could have on this analysis: completeness of attribute set and performance of ML method.

If an insufficient volume of data are labelled, the attribute set could be incomplete and not represent the situations sufficiently. However, in this case, it emerged that the set of 'frequent' attributes (those which occurred in greater than 1% of the labelled data) was not changing as more data were being labelled. This shows that saturation had been reached.

Additionally, as shown in the results of Chapter 5, training volume was a significant limitation in the achievement of high performance scores for the ML methods. If more labelled data were available, it is probably that high performances would be achieved. Future research and application of this method should consider the creation of larger labelled data sets.

Limited exploration of knowledge discovery methods

Chapter 6 presented three analysis methods for knowledge discovery (i.e. information generation) from attribute-based representations of failure events. These were included as use-cases to demonstrate the potential of these representations and the diverse information they could provide. While they succeeded in illustrating the advantages of attribute-based representations and facilitated discussion for application of these methods in organisational learning in construction, the exploration of the methods themselves and their results were deliberately truncated.

This truncation was due both to time and also a desire to remain focused on the development of the attribute-based methodology, rather than specific values obtained. For example, while this thesis could have explored at length the specific values of quantified risk presented in Section 6.2, it did not add to the topic of developing an approach to learning from text-based failure data. It would add to knowledge of safety failures in UK construction. However, there is worth in re-visiting the results of each method presented in Chapter 6 to exploit the possible knowledge gain in this additional area.

Additionally, these three are only a few of a multitude of analysis methods. Future investigations should examine a wider variety of analysis methods and in greater depth.

8.2.2 Future research themes

Data requirements and codification

The 'Golden Thread of Information' was introduced by Dame Judith Hackitt in 2018 in her final report recommendations following her independent review of building regulations and fire safety after the Grenfell Fire (Hackitt, 2018). She dedicates a whole chapter of the report to outlining the requirement for well-maintained, digital information which aligns with the information requirements of construction stakeholders - contractors, clients, owners and users. This is to "ensure that accurate building information is securely created, updated and accessible".

However, this research revealed that the data collected about failure events - especially regarding categorising events - was not always congruent with the epistemology of the information and analysis desired from them. Additionally, the data were often incomplete and accessibility was found to be challenging - recall that safety data were selected for investigation as other types of data were more difficult to access.

Additionally, the majority of data collection in construction has developed organically, based on what we *can* capture not what we *want* to capture. Current data collection processes should be reexamined in light of this change in perspective, especially considering the high level requirements of the process.

There is therefore much further work to be undertaken regarding the capture, curation and storage of data. Possible avenues for further investigation to develop the 'Golden Thread of Information' include:

- Exploration of suitable codification methods for explicit and tacit information - This research found that, to a certain extent, categorising failure events is a subjective task. Future research could explore how construction failure data containing different levels of bias should be collected and codified to discourage mis-analysis and misinterpretation of results. A possible method to explore could be replacing multiple choice categories with a series of simple, factual questions - similar to pre-screening questionnaires in the medical profession.
- Automation of data fields - A finding of this research found a conflict in the number of data fields required vs the time pressures on project. Future work should explore which fields can be automated and produce the same quality, or better quality, data. For example, weather data could be automated from archived weather, lost days for the injured person could be automated from timesheet data.
- Improvements to collection of remedial data - Remedial data, such as 'lessons learned' and 'suggestions to management' were particularly poor in the data explored here. Future research could explore barrier to capture/creation of this data and strategies to better capture this information. For example, can automated remedial 'checklists' be created based on the original data entried for the failure event and be sent following a specified time period/event stage (i.e. IP return to work or rework scheduled).

Future data types - video CCTV recordings, wearable tech and photos of the event - may also add a depth of information to the event representation. However, before including these data simply because we *can*, it should be considered what additional value they bring and how they can be analysed to bring that value into fruition. It could be noted that these are unstructured data types which would require translation to a structured data representation for inclusion in further analysis. In this way, they are also congruent with attribute-based representations. Attribute-based representations, which provide a level of anonymity, may also prove to be a method to deal with ethical considerations regarding these data sources.

Automating attribute list identification and creating taxonomies

Development of the attribute set for this analysis represents a significant proportion of the research presented. However, the current set and method is limited in its ability to adapt and update to new situations and contexts. Future research could consider more complex pipelines to automate or semi-automate the identification of the attribute set.

A possible avenue for investigation, outlined earlier, is the combination of NER-like identification of attribute types (i.e. actions, objects and site environment) and term clustering

A consideration for this investigation includes the preservation of anonymity in the attribute set. Any selected method cannot allow inclusion of specific names or places into the attribute set. This would significantly limit the usefulness of the data and could have ethical implications as well as possibly breaching GDPR.

Another consideration is that an 'industry standard' attribute set would facilitate greater analysis and benchmarking. In order to facilitate this, it is possible that rather than a flat list of attributes, it is more appropriate to consider a taxonomy. This is especially true when considering the applicability of these attributes to different failure criteria, e.g. analysis of safety and quality may quire different attribute granularities.

Creating taxonomy of attributes - which individual actions, objects and environment descriptors are grouped into categories and then split into more refined elements - would be a large undertaking. However, recent developments of BIM frameworks can significantly contribute in this space. BIM frameworks are extremely compatible with attribute-based analysis and Hallowell, Hardison, and Desvignes (2016) note that these technologies could support interoperability between different systems and data sources. This would bring huge impact to the construction industry.

Integrate other data types

Future research should also aim to integrate other data - for example, data for different failure criteria and project data - to exploit analysis for examination of interdependent features of construction.

A suggested way to integrate data is to have some common element/feature to connect the data. This common element could be a number of different features. For example, it could be the project, location, time, weather, contract type. It could also be common attribute, as defined by this research.

It is in this combination of data sets that truly unknown correlations and relationships will begin to emerge. For example, while attribute-based analysis alone could not identify root causes, we could add data such as programme (predicted and actual) to see whether other factors (e.g. time pressure) has an effect.

Action research and project trials

In Chapter 7 of this research, a number of suggestions were made for implementing this methodology into learning from failure processes in industry. However, these applications remain at a low technology readiness level and future research should look to undertake action research and site trials to ascertain the usefulness of different implementations of these knowledge discovery methods into project processes.

Increasing the performance scores of NLP+ML pipeline

This research achieved an F1 performance score of 54.2% for the best performing NLP+ML method (SVM) to predict attributes from unseen data. Human annotators achieved an agreement of 42.9%, which is equivalent to F1 = 66%. For deployment in industry, it is essential to increase the performance of the pipeline used so that trust is built in the process. Taking the human annotators as a gold standard, the pipeline currently operates at 75% of their value. As a rule of thumb, ML scientist speak of the ‘magical’ 90%, however, project trials and industry engagement should confirm this ‘target’ performance.

Future investigations should look to improve this performance. Some possible ways this required additional performance could be gained include: (a) more training data; (b) more complex pipelines which remove noise in the data (i.e. NER then classification) and (c) finer granularity of hyper-parameter and sampling optimisation.

8.3 Recommendations for application in industry

This research has resulted in several, succinct recommendations for immediate effect in industry. These are:

1. Appreciate that organisational learning is not (just) increasing human competence

   A common perception of the interviewees in Chapter 4 and members of industry I interacted with during the course of this research and my time working on construction sites was that learning was equivalent to developing human competence. While developing personnel is a key aspect of organisational learning, in this area, the whole is not a sum of its parts. The industry needs a better appreciation that continuous improvement processes and revision of processes/procedures also require attention from a learning perspective.

2. Systematise feedback / feed-forward

   Learning from failure depends upon efficient learning cycles, across different levels of the organisation which are trusted. The construction industry needs to examine how data and knowledge discovery tasks can be integrated into everyday processes to automatically transfer learning up and down the management hierarchy (feedback and feed-forward learning). This should move away from the generation and distribution of numerous alerts and develop continuous improvement processes which are less reliant on constant human intervention.

3. Digitise and centralise data

   This research struggled with inaccessibility of data. Several times, required data were found to be either not created/stored digitally to begin with or confined to disparate IT systems (or even worse, personal files!). A lack of digitisation and Centralised Data Environments (CDEs) plagues the construction industry, inhibiting the analysis, value and progress. These principles are essential to future-proof the industry and propel innovation and research effort, as well as creating efficiency savings for existing processes. This finding corroborates with the 'Golden Thread of Information', introduced by Dame Judith Hackitt in 2018 in her final report recommendations following her independent review of building regulations and fire safety after the Grenfell Fire (Hackitt, 2018).

4. Re-think *why* data is being collected

   It appears the majority of data collection in construction has developed organically, based on what we *can* capture not what we *want* to capture. Working in parallel to academia, industry should reexamine its current data collection processes in light of this change in perspective, especially considering the high level requirements of the process. This would lead to better quality, more useful data.

5. Develop more formal socialisation processes

   The construction industry is extremely dependent on knowledge transfer via socialisation. Formal processes, such as apprenticeships, target early-career, explicit, skills-based knowledge. This research recommends more investment in formal socialisation processes for tacit and through-life learning for all levels of worker. In particular, industry should invest in formal mentoring schemes for blue-collar workers, targeting career steering and so-called 'soft skills'.

6. Rethink lessons learned processes

   This research suggests that the primary aim of these exercises should be to maximise the tacit transfer of failure (and success) information, with capture of that information as a secondary priority. While the industry could also look to implement better methods of codification and retrieval of the lessons learned documents, using AI and smart search algorithms as in Eken et al. (2020), this does not address the root cause of why these databases are not used - people simply do not have time to use them.

8.4 Closing Remarks

This project harnessed the potential of modern data science methods, including natural language processing (NLP) and machine learning (ML), to produce automated methods and recommendations for analysing text-based failure data for the construction industry. A multi-method approach was applied.

First, a qualitative investigation used semi-structured interviews and thematic analysis to explore failure in the construction industry, with particular attention to present 'learning from failure' practice, human factors and biases.

Second, the text-based construction site failure data was analysed using recent data science methods. This analysis relied upon the insights from the first investigation to inform methodological decisions. It was decided to transform the unstructured text data into structured attributes, using machine learning classification methods, for further analysis. Transforming the unstructured text descriptions in this way allows further analysis methods to be performed. Possible further analyses unlocked by this method include risk analysis, graphical analysis, and finer trend analysis.

Finally, qualitative information from the thematic analysis was used to assess usefulness and form recommendations for industrial application of the data analysis methods employed to develop techniques that allow the capture and analysis of data to measure and mitigate the cumulative impact of smaller failures.

This is the first multi-method investigation into the use of text-based failure data for learning in the construction industry. This research contributed:

1. a greater depth to the understanding of the current state of learning from failure in construction;
2. a set of representation attributes for safety data in the UK. This used original data from a large UK construction company;
3. a Natural Language Processing (NLP) + Machine Learning (ML) pipeline using human-centre machine learning principles which can be trained to automatically extract attributes from text-based failure data in order to structure these data for further analysis;
4. a detailed cross-examination of the principles of this methodology and principles of general application of AI in construction; and
5. substantial recommendations for application of the findings into industry and avenues for future research.

Bibliography

A14 Improvement Scheme Progress (2020). URL: https://highwaysengland.co.uk/a14-cambridge-to-huntingdon-improvement-scheme-progress/.

A14 Integrated Delivery Team – A14 Cambridge to Huntingdon Improvement Scheme _ ccscheme (2019). URL: https://www.ccscheme.org.uk/ultrasite/a14-integrated-delivery-team-a14-cambridge-to-huntingdon-improvement-scheme/.

Abbasi, Ahmed, Brent Kitchens, and Faizan Ahmad (2019). *The Risks of AutoML and How to Avoid Them.* URL: https://hbr.org/2019/10/the-risks-of-automl-and-how-to-avoid-them.

Ahiaga-Dagbui, Dominic et al. (Oct. 2016). "Toward a Systemic View to Cost Overrun Causation in Infrastructure Projects: A Review and Implications for Research". In: *Project Management Journal.* ISSN: 1938-9507.

Akhavan, Peyman et al. (2016). "Major trends in knowledge management research: a bibliometric study". In: *Scientometrics* 107, pp. 1249–1264. DOI: 10.1007/s11192-016-1938-x.

Akhavian, Reza and Amir H. Behzadan (Oct. 2015). "Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers". In: *Advanced Engineering Informatics* 29.4, pp. 867–877. ISSN: 1474-0346. DOI: 10.1016/j.aei.2015.03.001. URL: https://www.sciencedirect.com/science/article/pii/S1474034615000282.

Al-Aubaidy, Nadia A, Carlos H Caldas, and Stephen P Mulva (2019). "Assessment of underreporting factors on construction safety incidents in US construction projects". In: *International Journal of Construction Management.* ISSN: 2331-2327. DOI: 10.1080/15623599.2019.1613211. URL: https://www.tandfonline.com/action/journalInformation?journalCode=tjcm20.

Al-Zwainy, Faiq M.S., Ibrahim A. Mohammed, and Ibrahim F. Varouqa (2018). "Diagnosing the Causes of Failure in the Construction Sector Using Root Cause Analysis Technique". In: *Journal of Engineering* 2018. ISSN: 23144912. DOI: 10.1155/2018/1804053.

Alruqi, Wael M and Matthew R Hallowell (2019). "Critical Success Factors for Construction Safety: Review and Meta-Analysis of Safety Leading Indicators". In: *Journal of Construction Engineering and Management* 145.3. DOI: 10.1061/(ASCE)CO.1943-7862.0001626.

Anders Örtenblad (2001). "On differences between organizational learning and learning organization". In: *The Learning Organization* 8.3, pp. 125–133. URL: http://www.emerald-library.com/ft.

Anumba, C. J. (Chimay J.), Charles O. Egbu, and Patricia M. Carrillo (2005). *Knowledge management in construction.* Blackwell Pub, p. 226. ISBN: 9781405129725.

Appelbaum, Steven H. "Socio-technical systems theory: an intervention strategy for organizational development". In: (). ISSN: 0021-1747.

Argote, L. (2011). "Organizational learning research: Past, present and future". In: *Management Learning* 42.4, pp. 439–446. ISSN: 1350-5076\n1461-7307. DOI: 10.1177/1350507611408217. URL: http://mlq.sagepub.com/cgi/doi/10.1177/1350507611408217.

Argyris, Chris (1977). “Organizational learning and management information systems”. In: *Accounting, Organizations and Society* 2.2, pp. 113–123.

Asrar-Ul-Haq, Muhammad and Sadia Anwar (2016). “A systematic review of knowledge management and knowledge sharing: Trends, issues, and challenges ”. In: *Cogent Business & Management* 3.1. ISSN: 2331-1975. DOI: 10.1080/23311975.2015.1127744. URL: https://doi.org/10.1080/23311975.2015.1127744.

Atkinson, Roger (1999). “Project management: Cost, time and quality, two best guesses and a phenomenon, its time to accept other success criteria”. In: *International Journal of Project Management* 17.6, pp. 337–342. ISSN: 02637863. DOI: 10.1016/S0263-7863(98)00069-6.

Baccarini, David (1999). “The Logical Framework Method for Defining Project Success”. In: *Project Management Journal* 30.4, pp. 25–32. DOI: 10.1177/875697289903000405.

Bacon, Francis (1869). “Novum Organum”. In: *The Works of Francis Bacon*. Ed. by J. Spedding, R.L. Ellis, and D.D. Heath. Vol. VIII. New York, 210–undefined.

Baker, Henrietta, Matthew R. Hallowell, and Antoine J.P. Tixier (Oct. 2020a). “AI-based prediction of independent construction safety outcomes from universal attributes”. In: *Automation in Construction* 118. ISSN: 09265805. DOI: 10.1016/j.autcon.2020.103146. URL: https://doi.org/10.1016/j.autcon.2020.103146.

— (Oct. 2020b). “Automatically learning construction injury precursors from text”. In: *Automation in Construction* 118. ISSN: 09265805. DOI: 10.1016/j.autcon.2020.103145. URL: https://doi.org/10.1016/j.autcon.2020.103145.

Bauer, F L and H Wossner (1972). “The "Plankalkul" of Konrad Zuse: A Fore-runner of Today’s Programming Languages”. In: *Communications of the ACM* 15.7, pp. 678–685.

Bengio, Yoshua et al. (2003). “A Neural Probabilistic Language Model Yoshua”. In: *Journal of Machine Learning Research* 3.Feb, pp. 1137–1155.

Bhatt, Umang et al. (Jan. 2020). “Explainable machine learning in deployment”. In: *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, pp. 648–657. ISBN: 9781450369367. DOI: 10.1145/3351095.3375624.

BIM maturity levels (2019). URL: https://www.designingbuildings.co.uk/wiki/BIM_maturity_levels.

Bishop, Christopher M (2006). *Pattern Recognition and Machine Learning*. Vol. 4. 4, p. 738. ISBN: 9780387310732. DOI: 10.1117/1.2819119. URL: http://www.library.wisc.edu/selectedtocs/bg0137.pdf.

Bojanowski, Piotr et al. (2017). “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Borgatti, Stephen P. and Brandon Ofem (2010). “Social network theory and analysis”. In: *Social network theory and educational change*, pp. 17–29.

Bottou, Léon, Frank E Curtis, and Jorge Nocedal (2018). “Optimization Methods for Large-Scale Machine Learning *”. In: *SIAM Review* 60.2, pp. 223–311. DOI: 10.1137/16M1080173. URL: http://www.siam.org/journals/ojsa.php.

Boysen, Philip G (2013). “Just Culture: A Foundation for Balanced Accountability and Patient Safety”. In: *The Ochsner* 13, pp. 400–406.

Breiman, Leo (1996a). “Bagging predictors”. In: *Machine learning* 24.2, pp. 123–140.

— (1996b). *Out-of-bag estimation.* Tech. rep. URL: https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf.

Breysse, Denys (2012). “Forensic engineering and collapse databases”. In: *Proceedings of the ICE - Forensic Engineering* 165.2, pp. 63–75. ISSN: 2043-9903. DOI: 10.1680/feng.10.00001. URL: http://www.icevirtuallibrary.com/content/article/10.1680/feng.10.00001.

BSI (2018). *PAS 1192-6: 2018: Specification for collaborative sharing and use of structured health and safety information using BIM.* Tech. rep.

Bye, Rolf Johan, Ragnar Rosness, and Jens Olgard Dalseth Røyrvik (2016). "'Culture' as a tool and stumbling block for learning: The function of 'culture' in communications from regulatory authorities in the Norwegian petroleum sector". In: *Safety Science* 81, pp. 68–80. ISSN: 18791042. DOI: 10.1016/j.ssci.2015.02.015. URL: http://dx.doi.org/10.1016/j.ssci.2015.02.015.

Caldas, Carlos H and Lucio Soibelman (2003). "Automating hierarchical document classification for construction management information systems". In: *Automation in Construction* 12.4, pp. 395–406. DOI: 10.1016/S0926-5805(03)00004-9.

Caldas, Carlos H et al. (2009). "Identification of Effective Management Practices and Technologies for Lessons Learned Programs in the Construction Industry". In: 135.June, pp. 531–539.

Campbell, Edward (2020). "Improving Construction Safety and Efficiency through a Method of Safety Risk Analysis". PhD thesis. University of Edinburgh.

Cannon, Mark D and Amy C Edmondson (2005). "Failing to Learn and Learning to Fail (Intelligently): How Great Organizations Put Failure to Work to Innovate and Improve". In: *Long Range Planning* 38, pp. 299–319. DOI: 10.1016/j.lrp.2005.04.005.

Castaneda, Delio Ignacio, Luisa Fernanda Manrique, and Sergio Cuellar (2018). "Is organizational learning being absorbed by knowledge management? A systematic review". In: *Journal of Knowledge Management* 22.2, pp. 299–325. ISSN: 17587484. DOI: 10.1108/JKM-01-2017-0041.

Chan, Albert P C, David Scott, and Edmond W M Lam (2002). "Framework of Success Criteria for Design/Build Projects". In: *Jounral of Management in Engineering* 18.3, pp. 120–128. DOI: 10.1061/ASCE0742-597X200218:3120.

Chan, Paul W and Robert C Moehler (2007). "Construction Skills Development in the UK: Transitioning Between the Formal and Informal". In: 2006, pp. 1–10. URL: http://ssrn.com/abstract=2141805.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* ACM, pp. 785–794.

Cheng, Min Yuan, Denny Kusoemo, and Richard Antoni Gosno (2020). "Text mining-based construction site accident classification using hybrid supervised machine learning". In: *Automation in Construction* 118.November 2019, p. 103265. ISSN: 09265805. DOI: 10.1016/j.autcon.2020.103265. URL: https://doi.org/10.1016/j.autcon.2020.103265.

Cheng, Min-Yuan et al. (2010). "Estimate at completion for construction projects using evolutionary support vector machine inference model". In: *Automation in Construction* 19.5, pp. 619–629. DOI: 10.1016/j.autcon.2010.02.008.

Chokor, Abbas et al. (2016). "Analyzing Arizona OSHA injury reports using unsupervised machine learning". In: *Procedia Engineering* 145, pp. 1588–1593.

Choo, Chun Wei (2000). "Working with knowledge: how information professionals help organisations manage what they know". In: *Library Management* 21.8, pp. 395–403. URL: http://www.mcbup.com/research_registers/lm.asp.

Chung, Sehwan (2018). "Bridge Damage Factor Recognition from Inspection Reports Using Active Recurrent Neural Network". PhD thesis. Seoul National University.

Claire, Iselin (2009). *Law Code of Hammurabi, king of Babylon.* URL: https://www.louvre.fr/en/oeuvre-notices/law-code-hammurabi-king-babylon.

Cole, John (2017). *Report of the Independent Inquiry into the Construction of Edinburgh Schools February 2017.* Tech. rep. February.

Costantino, Tracie (2008). *Constructivism.*

Currie, Wendy L. (Sept. 2012). "Institutional isomorphism and change: The national programme for IT - 10 years on". In: *Journal of Information Technology* 27.3, pp. 236–248. DOI: 10.1057/jit.2012.18.

Cyert, Richard and James Marsh (1963). "Behavioural theory of the firm". In: *Organisational Behaviour 2: Essential Theories of Process and Structure.* NJ: Englewood Cliffs. Chap. 4, pp. 169–187. ISBN: 0-7656-1525-8.

Daniel, Emmanuel Itodo et al. (2020). "Strategies for improving construction craftspeople apprenticeship training programme: Evidence from the UK". In: DOI: 10.1016/j.jclepro.2020.122135. URL: https://doi.org/10.1016/j.jclepro.2020.122135.

De Dianous, Valérie and Cécile Fiévez (2006). "ARAMIS project: A more explicit demonstration of risk control through the use of bow-tie diagrams and the evaluation of safety barrier performance". In: *Journal of Hazardous Materials* 130.3 SPEC. ISS. Pp. 220–233. ISSN: 03043894. DOI: 10.1016/j.jhazmat.2005.07.010.

Dekker, Sidney W A (2009). "Just culture: who gets to draw the line?" In: *Cognition, Technology & Work* 11, pp. 177–185. DOI: 10.1007/s10111-008-0110-7.

Delatte, Norbert (2010). "Failure literacy in structural engineering". In: *Engineering Structures* 32.7, pp. 1952–1954. ISSN: 01410296. DOI: 10.1016/j.engstruct.2009.12.015. URL: http://dx.doi.org/10.1016/j.engstruct.2009.12.015.

Desvignes, Matthieu (2014). "Requisite empirical risk data for integration of safety with advanced technologies and intelligent systems". PhD thesis. University of Colorado at Boulder.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805.*

Ding, Lieyun et al. (Feb. 2018). "A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory". In: *Automation in Construction* 86, pp. 118–124. ISSN: 09265805. DOI: 10.1016/j.autcon.2017.11.002.

Drupsteen, Linda and Frank W Guldenmund (2014). "What Is Learning ? A Review of the Safety Literature to Define Learning from Incidents , Accidents and Disasters". In: *Journal of Contingencies and Crisis Management* 22.2.

Drupsteen, Linda and Peter Hasle (2014). "Why do organizations not learn from incidents ? Bottlenecks , causes and conditions for a failure to effectively learn". In: *Accident Analysis and Prevention* 72, pp. 351–358. ISSN: 0001-4575. DOI: 10.1016/j.aap.2014.07.027. URL: http://dx.doi.org/10.1016/j.aap.2014.07.027.

Easterby-Smith, Mark, Mary Crossan, and Davide Nicolini (2000). "Organizational Learning: Debates Past, Present and Future". In: *Journal of Management Studies* 37.6. ISSN: 0022-2380.

Easterby-Smith, Mark and Marjorie A Lyles (2011). *The Evolving Field of Organizational Learning and Knowledge Management.* Tech. rep.

Easterby-Smith, Mark, Richard Thorpe, and Paul R. Jackson (2012). *Management research.* Sage.

Eken, Gorkem et al. (Feb. 2020). "A lessons-learned tool for organizational learning in construction". In: *Automation in Construction* 110. ISSN: 09265805. DOI: 10.1016/j.autcon.2019.102977.

Ekvall, Göran (Mar. 1996). "Organizational climate for creativity and innovation". In: *European Journal of Work and Organizational Psychology* 5.1, pp. 105–123. ISSN: 1359-432X. DOI: 10.1080/13594329608414845. URL: http://www.tandfonline.com/doi/abs/10.1080/13594329608414845.

EPIC (2016).

Esmaeili, Behzad (2012). "Identifying and quantifying construction safety risks at the attribute level". PhD thesis. University of Colorado, Boulder.

Esmaeili, Behzad, Matthew R Hallowell, and Balaji Rajagopalan (2015). "Attribute-based safety risk assessment. II: predicting safety outcomes using generalized linear models". In: *Journal of*

Construction Engineering and Management 141.8, p. 4015022. DOI: 10.1061/(ASCE)CO.1943-7862.0000981.

Esmi, Reza and Richard Ennals (2009). “Knowledge management in construction companies in the UK”. In: *AI and Society* 24.2, pp. 197–203. ISSN: 09515666. DOI: 10.1007/s00146-009-0202-9.

Fiol, C Marlene and Marjorie A Lyles (1985). “Organizational Learning”. In: *Academy of Management Review* 10.4, pp. 803–813. URL: https://www.jstor.org/stable/pdf/258048.pdf.

Fox, Nick (2008). *Postpositivism.* DOI: 10.4135/9781412963909.n332. URL: http://sk.sagepub.com.ezproxy.is.ed.ac.uk/reference/research/n332.xml.

Freund, Yoav and Robert E Schapire (1997). “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1, pp. 119–139.

Gammel, Josef et al. (2019). “A framework integrating technical, social, and managerial aspects of effective knowledge management”. In: *Proceedings of the European Conference on Knowledge Management, ECKM* 1, pp. 361–370. ISSN: 20488971. DOI: 10.34190/KM.19.106.

Gandhi, Prashant, Somesh Khanna, and Sree Ramaswamy (2016). *Which Industries Are the Most Digital (and Why)?* URL: https://hbr.org/2016/04/a-chart-that-shows-which-industries-are-the-most-digital-and-why.

Garvin, David A, Amy C Edmondson, and Francesca Gino (2008). “Is Yours a Learning Organization?” In: *Harvard Business Review* March. ISSN: 0017-8012.

Gergen, Kenneth J. and Mary M. Gergen (2008). *Social constructionism.*

Gillies, Marco et al. (May 2016). “Human-centered machine learning”. In: *Conference on Human Factors in Computing Systems - Proceedings.* Vol. 07-12-May-2016. Association for Computing Machinery, pp. 3558–3565. ISBN: 9781450340823. DOI: 10.1145/2851581.2856492.

Goh, Yang Miang and C. U. Ubeynarayana (2017). “Construction accident narrative classification: An evaluation of text mining techniques”. In: *Accident Analysis & Prevention* 108.May, pp. 122–130. ISSN: 00014575. DOI: 10.1016/j.aap.2017.08.026. URL: http://dx.doi.org/10.1016/j.aap.2017.08.026.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning.* MIT press.

Google (2000). *How Search algorithms work.* URL: https://www.google.com/search/howsearchworks/algorithms/.

Grimes, Seth (Aug. 2008). *Unstructured Data and the 80 Percent Rule.* Ed. by breakthroughanalysis.com.

Gursky, Jacob (2020). *Boosting Showdown_ Scikit-Learn vs XGBoost vs LightGBM vs CatBoost in Sentiment Classification.* URL: https://towardsdatascience.com/boosting-showdown-scikit-learn-vs-xgboost-vs-lightgbm-vs-catboost-in-sentiment-classification-f7c7f46fd956.

Hackitt, Judith (2018). *Building a Safer Future. Independent Review of Building Regulations and Fire Safety: Final Report.* Tech. rep. London. URL: www.gov.uk/government/publications.

Haixiang, Guo et al. (2017). *Learning from class-imbalanced data: Review of methods and applications.* DOI: 10.1016/j.eswa.2016.12.035.

Hallowell, Matthew R, Siddharth Bhandari, and Wael Alruqi (2019). “Methods of safety prediction: analysis and integration of risk assessment, leading indicators, precursor analysis, and safety climate”. In: *Construction Management and Economics*, pp. 1–14. DOI: 10.1080/01446193.2019.1598566.

Hallowell, Matthew R and John A Gambatese (2009a). “Activity-based safety risk quantification for concrete formwork construction”. In: *Journal of Construction Engineering and Management* 135.10, pp. 990–998.

Hallowell, Matthew R and John A Gambatese (2009b). “Construction Safety Risk Mitigation”. In: *Journal of Construction Engineering and Management* 135.12. DOI: 10.1061/ASCECO.1943-7862.0000107.

Hallowell, Matthew R et al. (2013). “Proactive Construction Safety Control: Measuring, Monitoring, and Responding to Safety Leading Indicators”. In: *Journal of Construction Engineering and Management* 139.10. DOI: 10.1061/(ASCE)CO.1943-7862.0000730. URL: https://ascelibrary-org.ezproxy.is.ed.ac.uk/doi/pdf/10.1061/%28ASCE%29CO.1943-7862.0000730.

Hallowell, Matthew Ryan, Dylan Hardison, and Matthieu Desvignes (July 2016). “Information technology and safety”. In: *Construction Innovation* 16.3, pp. 323–347. ISSN: 14770857. DOI: 10.1108/CI-09-2015-0047.

Hansen, M. T., N. Nohria, and T. Tierney (1999). “What’s your strategy for managing knowledge?” In: *Harvard business review* 77.2. ISSN: 00178012.

Harreveld, Bobby et al. (2016). *Constructing methodology for qualitative research: researching education and social practices.* Springer. ISBN: 978-1-137-59942-1.

Haslam, R. A. et al. (2005). “Contributing factors in construction accidents”. In: *Applied Ergonomics.* Vol. 36. 4 SPEC. ISS. Elsevier Ltd, pp. 401–415. DOI: 10.1016/j.apergo.2004.12.002.

Hastie, Trevor et al. (2005). “The elements of statistical learning: data mining, inference and prediction”. In: *The Mathematical Intelligencer* 27.2, pp. 83–85. DOI: 10.1007/BF02985802.

Heisig, Peter (July 2009). “Harmonisation of knowledge management – comparing 160 KM frameworks around the globe”. In: *Journal of Knowledge Management* 13.4, pp. 4–31. ISSN: 17587484. DOI: 10.1108/13673270910971798.

Helmreich, Robert L. (2000). “On error management: lessons from aviation”. In: *BJM* 320, pp. 781–785. DOI: 10.1136/bmj.320.7237.781.

Henke N., et al. et al. (2016). “The age of analytics: Competing in a data-driven world”. In: *McKinsey Global Institute* 4.December, p. 136. ISSN: 15543641. DOI: 10.1111/bjet.12230. URL: https://www.mckinsey.com/~/media/mckinsey/businessfunctions/mckinseyanalytics/ourinsights/theageofanalyticscompetinginadatadrivenworld/mgi-the-age-of-analytics-full-report.ashx.

Heraghty, Derek, Andrew J. Rae, and Sidney W.A. Dekker (2020). “Managing accidents using retributive justice mechanisms: When the just culture policy gets done to you”. In: *Safety Science* 126.February, p. 104677. ISSN: 18791042. DOI: 10.1016/j.ssci.2020.104677. URL: https://doi.org/10.1016/j.ssci.2020.104677.

Hey, Tony, Stewart Tansley, and Kristin Tolle (Oct. 2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research. ISBN: 978-0-9825442-0-4. URL: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

Hiles, David (2008). *Axiology.*

Hirschberg, Julia and Christopher D. Manning (2015). “Advances in natural language processing”. In: *Science* 349.6245, pp. 261–266.

Hoffmeister, Krista et al. (Sept. 2011). “A perspective on effective mentoring in the construction industry”. In: *Leadership and Organization Development Journal* 32.7, pp. 673–688. ISSN: 01437739. DOI: 10.1108/01437731111169997.

Hogarth, Terence and Lynn Gambin (2014). “Employer investment in Apprenticeships in England: an exploration of the sensitivity of employers in the construction sector to the net costs of training”. In: *Construction Management and Economics* 32.9, pp. 845–856. ISSN: 0144-6193.

DOI: 10.1080/01446193.2014.923577. URL: https://www.tandfonline.com/action/journalInformation?journalCode=rcme20.

Hoorn, Bronte van der and Stephen J. Whitty (May 2015). "A Heideggerian paradigm for project management: Breaking free of the disciplinary matrix and its Cartesian ontology". In: *International Journal of Project Management* 33.4, pp. 721–734. ISSN: 02637863. DOI: 10.1016/j.ijproman.2014.09.007.

HSE and Health and Safety Executive (2018). "Construction statistics in Great Britain, 2018". In: October, p. 20. URL: www.hse.gov.uk/statistics/http://www.hse.gov.uk/statistics/industry/construction.pdf.

Hubbard, R. K B and J. T. Neil (1985). "Major and minor accidents at the Thames Barrier construction site". In: *Journal of Occupational Accidents* 7.3, pp. 147–164. ISSN: 03766349. DOI: 10.1016/0376-6349(85)90001-X.

ICE (2019). *Reducing the gap between cost estimates and outturns for major infrastructure projects and programmes.* Tech. rep.

Inkinen, Henri (Apr. 2016). *Review of empirical research on knowledge management practices and firm performance.* DOI: 10.1108/JKM-09-2015-0336.

Johnston, Alan (2012). "Rigour in research : theory in the research approach". In: DOI: 10.1108/EBR-09-2013-0115.

Jones, Karen Spärck (2004). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 60, pp. 11–21. ISSN: 0022-0418.

Kang, Kyungsu and Hanguk Ryu (2019). "Predicting types of occupational accidents at construction sites in \uppercase{K}orea using random forest model". In: *Safety Science* 120, pp. 226–236. ISSN: 18791042. DOI: 10.1016/j.ssci.2019.06.034.

Katz, Slava (1987). "Estimation of probabilities from sparse data for the language model component of a speech recognizer". In: *IEEE transactions on acoustics, speech, and signal processing* 35.3, pp. 400–401.

Kim, Hyun-Chul et al. (2002). "Support Vector Machine Ensemble with Bagging". In: *Pattern Recognition with Support Vector Machines.* Ed. by Seong-Whan Lee and Alessandro Verri. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 397–408. ISBN: 978-3-540-45665-0. DOI: 10.1007/3-540-45665-1{_}31.

Kim, Taekhyung and Seokho Chi (2019). "Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry". In: *Journal of Construction Engineering and Management* 145.3, p. 4019004. ISSN: 0733-9364. DOI: 10.1061/(asce)co.1943-7862.0001625.

Kim, Yoon (2014). "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882.* ISSN: 15206149. DOI: 10.1145/1599272.1599278. URL: http://arxiv.org/abs/1408.5882.

Koehn, Enno and Kurt Musser (1983). "OSHA regulations effects on construction". In: *Journal of Construction Engineering and Management* 109.2, pp. 233–244. DOI: 10.1061/(ASCE)0733-9364(1983)109:2(233).

Kohn, Linda T., Janet M. Corrigan, and Molla S. Donaldson (2000). *To Err Is Human. Building a Safer Health System, Volume 6.* Vol. 2. 3, pp. 93–95. ISBN: 0309261740. DOI: 10.17226/9728. URL: https://books.google.com/books?hl=en&lr=&id=Jj25GlLKXSgC&pgis=1.

Kolb, David A (2015). *Experiential Learning: Experience as The Source of Learning and Development.* Second Edition, pp. 20–38. ISBN: 0017-8012. DOI: 10.1016/B978-0-7506-7223-8.50017-4.

Kululanga, G. K., A. D. F. Price, and R. McCaffer (2002). "Empirical Investigation of Construction Contractors' Organizational Learning". In: *Journal of Construction Engineering and*

Management 128.5, pp. 385–391. ISSN: 0733-9364. DOI: 10.1061/(asce)0733-9364(2002)128:5(385).

Lam, Ka Chi, Ekambaram Palaneeswaran, and Chen-yun Yu (2009). "A support vector machine model for contractor prequalification". In: *Automation in Construction* 18.3, pp. 321–329. DOI: 10.1016/j.autcon.2008.09.007.

Lampel, Joseph et al. (Sept. 2009). "Experiencing the improbable: Rare events and organizational learning". In: *Organization Science* 20.5, pp. 835–845. ISSN: 10477039. DOI: 10.1287/orsc.1090.0479.

Le, James (2018). *k-Nearest Neighbors: Who are close to you?* URL: https://medium.com/cracking-the-data-science-interview/k-nearest-neighbors-who-are-close-to-you-19df59b97e7d.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Levitt, Barbara and James G. March (1988). "Organizational Learning". In: *Annual Review of Sociology* 141.1, pp. 319–338. ISSN: 0201001748.

Li, Bing et al. (June 2005). "Critical success factors for PPP/PFI projects in the UK construction industry". In: *Construction Management and Economics* 23.5, pp. 459–471. ISSN: 01446193. DOI: 10.1080/01446190500041537.

Lin, Gongbo and Qiping Shen (2007). "Measuring the Performance of Value Management Studies in Construction: Critical Review". In: *Journal of Management in Engineering* 23.1, pp. 2–9. DOI: 10.1061/ASCE0742-597X200723:12.

Littlejohn, Allison et al. (2017). "Learning from Incidents Questionnaire (LFIQ): The validation of an instrument designed to measure the quality of learning from incidents in organisations". In: *Safety Science* 99, pp. 80–93. ISSN: 18791042. DOI: 10.1016/j.ssci.2017.02.005. URL: http://dx.doi.org/10.1016/j.ssci.2017.02.005.

Lloyd-walker, Beverley M et al. (2014). "Enabling construction innovation : the role of a no-blame culture as a collaboration behavioural driver in project alliances Enabling construction innovation : the role of a no-blame culture as a collaboration behavioural driver in project alliances". In: *Construction Management and Economics* 32.3, pp. 229–245. ISSN: 0144-6193. DOI: 10.1080/01446193.2014.892629. URL: http://dx.doi.org/10.1080/01446193.2014.892629.

Love, Peter and Jim Smith (2019). "Unpacking the ambiguity of rework in construction: making sense of the literature". In: *Civil Engineering and Environmental Systems* 35.1-4, pp. 180–203. DOI: 10.1080/10286608.2019.1577396. URL: https://iahr.tandfonline.com/doi/pdf/10.1080/10286608.2019.1577396?needAccess=true.

Love, Peter E.D. et al. (2015). "The symbiotic nature of safety and quality in construction: Incidents and rework non-conformances". In: *Safety Science*. ISSN: 18791042. DOI: 10.1016/j.ssci.2015.05.009.

Luft, Joan and Michael D. Shields (Oct. 2014). "Subjectivity in developing and validating causal explanations in positivist accounting research". In: *Accounting, Organizations and Society* 39.7, pp. 550–558. ISSN: 03613682. DOI: 10.1016/j.aos.2013.09.001. URL: http://dx.doi.org/10.1016/j.aos.2013.09.001.

Lukic, Dane, Allison Littlejohn, and Anoush Margaryan (2012). "A framework for learning from incidents in the workplace". In: *Safety Science* 50.4, pp. 950–957. ISSN: 0925-7535. DOI: 10.1016/j.ssci.2011.12.032. URL: http://dx.doi.org/10.1016/j.ssci.2011.12.032.

Lukic, Dane, Anoush Margaryan, and Allison Littlejohn (2013). "Individual agency in learning from incidents". In: *Human Resource Development International* 16.4, pp. 409–425. ISSN:

1469-8374. DOI: 10.1080/13678868.2013.792490. URL: https://www.tandfonline.com/action/journalInformation?journalCode=rhrd20.

Lundberg, Mary, Helena Lidelöw, and Susanne Engström (2017). “Methods used for knowledge management and organizational learning in the practice of construction projects: A systematic literature review”. In: *Proceedings of working papers from the ARCOM and BEAM Centre Early Career Researcher and Doctoral Workshop on Building Asset Management*. Ed. by Craig Thomson, pp. 30–40. URL: www.gcu.ac.uk/assetmanagement/.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning (2015). “Effective approaches to attention-based neural machine translation”. In: *arXiv preprint arXiv:1508.04025*.

Margaryan, Anoush, Allison Littlejohn, and Neville A Stanton (2017). “Research and development agenda for Learning from Incidents”. In: *Safety Science* 99.A, pp. 5–13. DOI: 10.1016/j.ssci.2016.09.004. URL: http://dx.doi.org/doi:10.1016/j.ssci.2016.09.004.

Marsick, Victoria and K.E. Watkins (2003). “Demonstrating the value of a Organizational Learning Culture: The Dimensions of the Learning Organization”. In: *Advances in Developing Human Resources* 5.2, pp. 132–151. ISSN: 15234223. DOI: 10.1177/1523422303251341.

Marzouk, Mohamed and Mohamed Enaba (2019). “Text analytics to analyze and monitor construction project contract and correspondence”. In: *Automation in Construction* 98.November 2018, pp. 265–274. ISSN: 09265805. DOI: 10.1016/j.autcon.2018.11.018. URL: https://doi.org/10.1016/j.autcon.2018.11.018.

McCaslin, Mark (2008). *Pragmatism.*

Mikolov, Tomas et al. (2013a). “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*, pp. 3111–3119.

Mikolov, Tomas et al. (2013b). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*. URL: http://ronan.collobert.com/senna/.

Miller, Robert L. and John Brewer (2003). *Attitude.* DOI: https://dx.doi.org/10.4135/9780857020024. URL: https://methods.sagepub.com/reference/the-a-z-of-social-research/n4.xml?fromsearch=true.

Moon, Soenghyeon Seonghyeon et al. (2018). “Analysis of Construction Accidents Based on Semantic Search and Natural Language Processing”. In: *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction.* Vol. 35. Isarc. IAARC Publications, pp. 1–6. DOI: 10.22260/isarc2018/0109.

Morland, Kate V, Dermot Breslin, and Fionn Stevenson (2019). “Development of a multi-level learning framework”. In: *The Learning Organisation* 26.1, pp. 78–96. DOI: 10.1108/TLO-04-2018-0080. URL: www.emeraldinsight.com/0969-6474.htm.

Moselhi, Osama, Tarek Hegazy, and Paul Fazio (1991). “Neural networks as tools in construction”. In: *Journal of Construction Engineering and Management* 117.4, pp. 606–625. DOI: 10.1061/(ASCE)0733-9364(1991)117:4(606).

Munier, Nolberto (2016). *Risk management for engineering projects.* Springer International. ISBN: 9783319052519. DOI: 10.1007/978-3-319-05251-9. URL: https://link.springer.com/book/10.1007%2F978-3-319-05251-9.

Murray, Barney (2020). “Named Entity Recognition for Construction Injury Reports”. PhD thesis.

Nath, N and Amir H Behzadan (2017). “Construction productivity and ergonomic assessment using mobile sensors and machine learning”. In: *Proceedings of the ASCE International Workshop on Computing in Civil Engineering 2017: Smart Safety, Sustainability and Resilience, Seattle, WA*, pp. 434–441. DOI: 10.1061/9780784480847.054.

Nikolova, Irina et al. (2014). “Learning climate scale: Construction, reliability and initial validity evidence”. In: *Journal of Vocational Behavior* 85.3, pp. 258–265. ISSN: 00018791. DOI: 10.1016/j.jvb.2014.07.007. URL: http://dx.doi.org/10.1016/j.jvb.2014.07.007.

Nonaka, Ikujiro (1991). "The knowledge-creating company". In: *Harvard Business Review*.

Nordhaus, William D (2001). *The Progress of Computing*. Tech. rep. Yale University. URL: http://cowles.econ.yale.edu/.

Nowak, Stefanie and Stefan Rüger (2010). "How reliable are annotations via crowdsourcing? a study about inter-annotator agreement for multi-label image annotation Conference or Workshop Item How Reliable are Annotations via Crowdsourcing? A Study about Inter-annotator Agreement for Multi-label Image Annotation". In: DOI: 10.1145/1743384.1743478. URL: http://dx.doi.org/doi:10.1145/1743384.1743478.

Office of National Statistics (2018). "Construction Industry: Statistics and policy". In: *House of Commons Library* 01432, pp. 1–13.

Osei-Kyei, Robert and Albert P C Chan (2017). "Comparative Analysis of the Success Criteria for Public–Private Partnership Projects in Ghana and Hong Kong". In: *Project Management Journal* 48.4, pp. 80–92.

Oswald, David (2020). "Construction Management and Economics Safety indicators: questioning the quantitative dominance". In: *Construction Management and Economics* 38.1, pp. 11–17. ISSN: 0144-6193. DOI: 10.1080/01446193.2019.1605184. URL: https://www.tandfonline.com/action/journalInformation?journalCode=rcme20.

Oswald, David et al. (2018). "An exploration into the implications of the 'compensation culture' on construction safety". In: *Safety Science* 109.May, pp. 294–302. ISSN: 18791042. DOI: 10.1016/j.ssci.2018.06.009. URL: https://doi.org/10.1016/j.ssci.2018.06.009.

Ouriques, Raquel Andrade Barros et al. (2018). "KNOWLEDGE MANAGEMENT STRATEGIES AND PROCESSES IN AGILE SOFTWARE DEVELOPMENT: A SYSTEMATIC LITERATURE REVIEW". In: *arXiv*.

Paley, John (2008). *Positivism*. DOI: doi:10.4135/9781412963909.n329.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Perrow, Charles. (1999). *Normal accidents : living with high-risk technologies*. Princeton University Press, p. 451. ISBN: 0691004129.

Peters, Matthew E et al. (2018). "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365*.

Pfatteicher, Sarah K A and D Ph (2000). "Walkways : Tragedy and Transformation in Kansas City". In:

Pinto, Jeffrey K and Samuel J Mantel (1990). *The Causes of Project Failure*. Tech. rep. 4.

Poh, Clive Q.X. X, Chalani Udhyami Ubeynarayana, and Yang Miang Goh (2018). "Safety leading indicators for construction sites: a machine learning approach". In: *Automation in Construction* 93, pp. 375–386. ISSN: 09265805. DOI: 10.1016/j.autcon.2018.03.022.

Pollack, Julien, Jane Helm, and Daniel Adler (2018). "What is the Iron Triangle, and how has it changed?" In: *International Journal of Managing Projects in Business* 11.2, pp. 527–547. ISSN: 17538386. DOI: 10.1108/IJMPB-09-2017-0107.

Poovey, Mary (1998). *A history of the modern fact*. Chicago: University of Chicago Press. ISBN: 0226675254.

Prolegomenon. URL: https://www.merriam-webster.com/dictionary/prolegomenon.

RAEng, The Royal Academy of Engineering (2010). *Philosophy of Engineering Volume 1 of the proceedings of a series of seminars held at The Royal Academy of Engineering*. Tech. rep. URL: www.raeng.org.uk/philosophyofengineering.

Reason, James (1990). “The contribution of latent human failures to the breakdown of complex systems”. In: *Trans. R. Soc. Lond. B* 327, pp. 475–484.

— (1998). “Achieving a safe culture: Theory and practice”. In: *Work & Stress* 12.3, pp. 293–306. ISSN: 1464-5335. DOI: `10.1080/02678379808256868`. URL: `https://doi.org/10.1080/02678379808256868`.

— (2000). “Education and debate Human error: models and management”. In: *BMJ* 320, pp. 768–770. DOI: `10.1136/bmj.320.7237.768`. URL: `http://www.bmj.com/`.

Reinsel, David, John Gantz, and John Rydning (2018). “Data Age 2025: The Digitization of the World From Edge to Core”. In: *Idc* November. URL: `https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf`.

Ren, Xu, Xiaofang Deng, and Lihua Liang (2018). “Knowledge transfer between projects within project-based organizations: the project nature perspective”. In: *Journal of Knowledge Management* 22.5, pp. 1082–1103. ISSN: 17587484. DOI: `10.1108/JKM-05-2017-0184`.

Richardson, M (2004). *Hammurabi's Laws: Text, Translation and Glossary.* T&T Clark International. URL: `https://books.google.co.uk/books?hl=en&lr=&id=UgfUAwAAQBAJ&oi=fnd&pg=PA5&dq=hammurabi%27s+code+of+laws+&ots=LeANsGri16&sig=MNBhd914gYqQGiOBA0Y01kLj3HU&redir_esc=y#v=onepage&q=hammurabi's%20code%20of%20laws&f=false`.

Rogers, Everett M (2003). *Diffusion of Innovations.* 5th.

Ruijter, A. de and F. Guldenmund (Apr. 2015). “The bowtie method: A review”. In: *Safety Science* 88, pp. 211–218. ISSN: 18791042. DOI: `10.1016/j.ssci.2016.03.001`.

Salminen, Simo (1995). “Serious occupational accidents in the construction industry”. In: *Construction Management and Economics* 13.4, pp. 299–306. DOI: `10.1080/01446199500000035`.

Samuel Smiles (1856). “Self-Culture: Facilities and Difficulties”. In: *Self-Help.* Chap. 11.

Sanders, Robert (1987). “THE PARETO PRINCIPLE: ITS USE AND ABUSE”. In: *The Journal of Services Marketing* 1.2.

Sanne, Johan M (2008). “Incident reporting or storytelling? Competing schemes in a safety-critical and hazardous work setting”. In: *Safety Science* 46, pp. 1205–1222. DOI: `10.1016/j.ssci.2007.06.024`.

Saqib, Muhammad, Rizwan U Farooqui, and Sarosh Lodi (2010). *Assessment of critical success factors for construction projects in pakistan Earthquake Model For Middle East Reigon View project Improving construction safety practices in Pakistan View project.* Tech. rep. URL: `https://www.researchgate.net/publication/292767223`.

Sarkar, Sobhan et al. (2019a). “An optimization-based decision tree approach for predicting slip-trip-fall accidents at work”. In: *Safety Science* 118, pp. 57–69. ISSN: 18791042. DOI: `10.1016/j.ssci.2019.05.009`.

Sarkar, Sobhan et al. (2019b). “Application of optimized machine learning techniques for prediction of occupational accidents”. In: *Computers & Operations Research* 106, pp. 210–224. DOI: `10.1016/j.cor.2018.02.021`. URL: `https://doi.org/10.1016/j.cor.2018.02.021`.

Saunders, M., P. Lewis, and A. Thornhill (2009). *Research Methods for Business Students.* 5th Edition. Harlow: Prentices-Hall.

Schmidhuber, Jürgen (2015). “Deep learning in neural networks: An overview”. In: *Neural networks* 61, pp. 85–117.

Shang, Guokan et al. (2019). “Energy-based Self-attentive Learning of Abstractive Communities for Spoken Language Understanding”. In: *arXiv preprint arXiv:1904.09491.*

Shenhar, Aaron J. et al. (2001). “Project success: A multidimensional strategic concept”. In: *Long Range Planning* 34, pp. 699–725. ISSN: 00246301. DOI: `10.1016/S0024-6301(01)00097-8`.

Silva, Sílvia A et al. (2017). “Organizational practices for learning with work accidents throughout their information cycle”. In: 99, pp. 102–114. DOI: `10.1016/j.ssci.2016.12.016`.

Silverman, David (2013). *A Very Short, Fairly Interesting and Reasonably Cheap Book about Qualitative Research*. 2nd. London: Sage. URL: https://books.google.co.uk/books?hl=en&lr=&id=hYcQAgAAQBAJ&oi=fnd&pg=PP2&dq=silverman+2007+methodology+book&ots=OSnN3O4_3a&sig=IjoZXg2SL1bZPuEPJTFZ62CMjms#v=onepage&q=silverman2007methodologybook&f=false.

Singhal, Abhay (2019). *AutoML: Opportunities and Challenges*. URL: https://medium.com/datadriveninvestor/automl-3803e315d5cd.

Skibniewski, MirosŁaw, Tomasz Arciszewski, and Kamolwan Lueprasert (1997). "Constructability analysis: machine learning approach". In: *Journal of Computing in Civil Engineering* 11.1, pp. 8–16. DOI: 10.1061/(ASCE)0887-3801(1997)11:1(8).

Smith, Simon David, Fred Sherratt, and David Oswald (2017). "The antecedents and development of unsafety". In: *Management, Procurement and Law* 170.MP2, pp. 59–67.

Soane, Alastair (2016). "Learning from experience to avoid collapse". In: *Proceeding of the Institution of Civil Engineers: Forensic Engineering* 169.4, pp. 127–132. ISSN: 20439911. DOI: 10.1680/jfoen.16.00004. URL: http://www.icevirtuallibrary.com/doi/pdf/10.1680/jfoen.16.00004.

Soibelman, Lucio and Hyunjoo Kim (2002). "Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases". In: *Journal of Computing in Civil Engineering* 16.1, pp. 39–48. ISSN: 08873801. DOI: 10.1061/(ASCE)0887-3801(2002)16:1(39).

Son, Hyojoo, Changmin Kim, and Changwan Kim (2011). "Automated color model–based concrete detection in construction-site images by using machine learning algorithms". In: *Journal of Computing in Civil Engineering* 26.3, pp. 421–433. DOI: 10.1061/(ASCE)CP.1943-5487.0000141.

Sooyoung, Choe and Leite Fernanda (2016). "Assessing Safety Risk among Different Construction Trades: Quantitative Approach". In: *Journal of Construction Engineering and Management* 143.5. DOI: 10.1061/(ASCE)CO.1943-7862.0001237.

Stemn, Eric et al. (2018). "Failure to learn from safety incidents : Status , challenges and opportunities". In: 101.August 2017, pp. 313–325. DOI: 10.1016/j.ssci.2017.09.018.

Structure. URL: https://languages.oup.com/google-dictionary-en/.

Sun, Jun et al. (2020). "Text visualization for construction document information management". In: *Automation in Construction* 111.2020, pp. 1–12. ISSN: 09265805. DOI: 10.1016/j.autcon.2019.103048.

Syed, Matthew (2015). *Black Box Thinking*. London: John Murray (Publishers). ISBN: 9781473613799.

Tabish, S. Z.S. and Kumar Neeraj Jha (Aug. 2011). *Identification and evaluation of success factors for public construction projects*. DOI: 10.1080/01446193.2011.611152.

Thompson, Paul et al. (2020). "Semantic Annotation for Improved Safety in Construction Work". In: *Proceedings of the 12th Conference on Language Resources and Evaluation*. Marseille, pp. 1990–1999. URL: http://www.nactem.ac..

Tixier, Antoine J-P, Matthew R Hallowell, and Balaji Rajagopalan (2017). "Construction safety risk modeling and simulation". In: *Risk analysis* 37.10, pp. 1917–1935.

Tixier, Antoine J-P et al. (2016a). "Application of machine learning to construction injury prediction". In: 69, pp. 102–114.

— (2016b). "Automated content analysis for construction safety : A natural language processing system to extract precursors and outcomes from unstructured injury reports". In: *Automation in Construction* 62, pp. 45–56. ISSN: 0926-5805. DOI: 10.1016/j.autcon.2015.11.001. URL: http://dx.doi.org/10.1016/j.autcon.2015.11.001.

Tixier, Antoine J-P P., Michalis Vazirgiannis, and Matthew R. Hallowell (2016). "Word Embeddings for the Construction Domain". In: *arXiv preprint arXiv:1610.09333*. URL: http://arxiv.org/abs/1610.09333.

Toor, Shamas ur Rehman and Stephen O. Ogunlana (May 2008). "Critical COMs of success in large-scale construction projects: Evidence from Thailand construction industry". In: *International Journal of Project Management* 26.4, pp. 420–430. ISSN: 02637863. DOI: 10.1016/j.ijproman.2007.08.003.

Tripathi, K. K. and K. N. Jha (Apr. 2018). "An Empirical Study on Performance Measurement Factors for Construction Organizations". In: *KSCE Journal of Civil Engineering* 22.4, pp. 1052–1066. ISSN: 19763808. DOI: 10.1007/s12205-017-1892-z.

Tuomi, Ilkka (1999). "Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory". In: *Journal of Management Information Systems* 16.3, pp. 103–117. DOI: 10.1080/07421222.1999.11518258. URL: https://www.tandfonline.com/doi/pdf/10.1080/07421222.1999.11518258?needAccess=true.

UK BIM Alliance (2019). *Information management according to BS EN ISO 19650. Guidance Part 1: Concepts.* Tech. rep.

Van Der Hoorn, Bronte and Stephen J. Whitty (Aug. 2015). "Signs to dogma: A Heideggerian view of how artefacts distort the project world". In: *International Journal of Project Management* 33.6, pp. 1206–1219. ISSN: 02637863. DOI: 10.1016/j.ijproman.2015.02.011.

Walker, Derek H.T. (Apr. 2016). "Reflecting on 10 years of focus on innovation, organisational learning and knowledge management literature in a construction project management context". In: *Construction Innovation* 16.2, pp. 114–126. ISSN: 14770857. DOI: 10.1108/CI-12-2015-0066.

Wanberg, John et al. (2013). "Relationship between Construction Safety and Quality Performance". In: *Journal of Construction Engineering and Management* 139.10. DOI: 10.1061/(ASCE). URL: https://ascelibrary-org.ezproxy.is.ed.ac.uk/doi/pdf/10.1061/%28ASCE%29CO.1943-7862.0000732.

Wang, Hao and Xianhai Meng (May 2019). *Transformation from IT-based knowledge management into BIM-supported knowledge management: A literature review.* DOI: 10.1016/j.eswa.2018.12.017.

Williams, Trefor P and Jie Gong (2014). "Predicting construction cost overruns using text mining, numerical data and ensemble classifiers". In: *Automation in Construction* 43, pp. 23–29. ISSN: 0926-5805. DOI: 10.1016/j.autcon.2014.02.014. URL: http://dx.doi.org/10.1016/j.autcon.2014.02.014.

Winch, Christopher and Linda Clarke (2010). "Oxford Review of Education 'Front-loaded' Vocational Education versus Lifelong Learning. A Critique of Current UK Government Policy". In: ISSN: 1465-3915. DOI: 10.1080/0305498032000080701. URL: https://www.tandfonline.com/action/journalInformation?journalCode=core20.

Woods, David D, Emily S Patterson, and Emilie M Roth (2002). "Can we ever escape from data overload? A cognitive systems diagnosis". In: *Cognition, Technology & Work* 4.1, pp. 22–36.

Yang, Huan et al. (Sept. 2010). "A critical review of performance measurement in construction". In: *Journal of Facilities Management* 8.4, pp. 269–284. ISSN: 17410983. DOI: 10.1108/14725961011078981.

Yao, Mariya (2017). *4 Unique Challenges Of Industrial Artificial Intelligence.* URL: https://www.forbes.com/sites/mariyayao/2017/04/14/unique-challenges-of-industrial-artificial-intelligence-general-electric/?sh=6317d8271305#59e377551305.

Yong, Yee Cheong and Nur Emma Mustaffa (Sept. 2013). "Critical success factors for Malaysian construction projects: An empirical assessment". In: *Construction Management and Economics* 31.9, pp. 959–978. ISSN: 01446193. DOI: 10.1080/01446193.2013.828843.

Yu, Wen-der and Jia-yang Hsu (2013). "Content-based text mining technique for retrieval of CAD documents". In: *Automation in construction* 31, pp. 65–74. ISSN: 0926-5805. DOI: 10.1016/j.autcon.2012.11.037. URL: http://dx.doi.org/10.1016/j.autcon.2012.11.037.

Zhang, Fan et al. (2019). "Construction site accident analysis using text mining and natural language processing techniques". In: *Automation in Construction* 99.June 2018, pp. 238–248. ISSN: 09265805. DOI: 10.1016/j.autcon.2018.12.016. URL: https://doi.org/10.1016/j.autcon.2018.12.016.

Zhang, Sijie, Frank Boukamp, and Jochen Teizer (2015). "Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (JHA)". In: *Automation in Construction* 52. ISSN: 09265805. DOI: 10.1016/j.autcon.2015.02.005.

Zhang, Xiang, Junbo Zhao, and Yann LeCun (2015). "Character-level convolutional networks for text classification". In: *Advances in neural information processing systems*, pp. 649–657.

Zhong, Botao et al. (2020). "Deep learning and network analysis: Classifying and visualizing accident narratives in construction". In: *Automation in Construction* 113.January, p. 103089. ISSN: 09265805. DOI: 10.1016/j.autcon.2020.103089. URL: https://doi.org/10.1016/j.autcon.2020.103089.

Zhou, Zhipeng, Yang Miang Goh, and Qiming Li (2015). "Overview and analysis of safety management studies in the construction industry". In: *Safety Science* 72, pp. 337–350. ISSN: 18791042. DOI: 10.1016/j.ssci.2014.10.006. URL: http://dx.doi.org/10.1016/j.ssci.2014.10.006.

Zou, Yang, Arto Kiviniemi, and Stephen W. Jones (2017). "Retrieving similar cases for construction project risk management using Natural Language Processing techniques". In: *Automation in Construction* 80.April, pp. 66–76. ISSN: 09265805. DOI: 10.1016/j.autcon.2017.04.003. URL: http://dx.doi.org/10.1016/j.autcon.2017.04.003.

Appendix A

Ethics and Data Management

A.1 Ethical Consent for Interviews

THE UNIVERSITY
of EDINBURGH

INTERVIEW CONSENT FOR SEMI-STRUCTURED INTERVIEW

PROJECT TITLE: LEARNING FROM FAILURE: A STUDY OF ATTITUDES TO LEARNING IN THE CONSTRUCTION INDUSTRY

PRINCIPAL INVESTIGATOR(S): HENRIETTA BAKER, DR SIMON SMITH

The interview will take 45 minutes. We do not anticipate that there are any risks associated with your participation, but you have the right to stop the interview or withdraw from the research at any time.

Thank you for agreeing to be interviewed as part of the above research project. Ethical procedures for academic research undertaken from UK institutions require that interviewees explicitly agree to being interviewed and how the information contained in their interview will be used. This consent form is necessary for us to ensure that you understand the purpose of your involvement and that you agree to the conditions of your participation. Would you therefore read and then sign this form to certify that you approve the following:

- the interview will be recorded and a transcript will be produced

- the transcript of the interview will be analysed by Henrietta Baker as research investigator

- access to the interview transcript will be limited to Henrietta Baker and academic colleagues and researchers with whom he might collaborate as part of the research process

- any summary interview content, or direct quotations from the interview, that are made available through academic publication or other academic outlets will be anonymised so that you cannot be identified, and care will be taken to ensure that other information in the interview that could identify yourself is not revealed

- any variation of the conditions above will only occur with your further explicit approval

If you wish to review the notes, transcripts, or other data collected during the research pertaining to my participation, please indicate so here:

	I wish to review the notes, transcripts, or other data collected during the research pertaining to my participation.

All or part of the content of your interview may be used;

- In academic or conference papers
- In other media that we may produce such as spoken presentations
- On other feedback events
- In an archive of the project as noted above

By signing this form I agree that;

1. I am voluntarily taking part in this project. I understand that I don't have to take part, and I can stop the interview at any time;
2. The transcribed interview or extracts from it may be used as described above;
3. I don't expect to receive any benefit or payment for my participation;
4. I can request a copy of the transcript of my interview and may make edits I feel necessary to ensure the effectiveness of any agreement made about confidentiality;
5. I have been able to ask any questions I might have, and I understand that I am free to contact the researcher with any questions I may have in the future.

____

Printed Name

____________________________________ ____________________

Participants Signature **Date**

____________________________________ ____________________

Researchers Signature **Date**

Contact Information

If you have any further questions or concerns about this study, please contact:
Name of researcher: Henrietta Baker
E-mail: s1679725@ed.ac.uk
You can also contact the research supervisor:
Name of researcher: Dr Simon Smith
E-mail: simon.smith@ed.ac.uk

A.2 Data Management Plan

PhD Construction Engineering - Facilitating Learning from Failure through Extracting Insights from Construction Site Free-Text Data

Data Collection

What data will you collect or create?

Interview data - digital voice recordings and notes from semi-structured interviews with members of the construction industry. These interviews will revolve around the notion of failure during construction and how data and learning is currently gleaned from these events.
Safety record data - safety records of accidents and near misses (unsafe conditions/actions) on construction sites.
Quality records - quality records of non-compliant, unacceptable or 'snag' quality events on construction sites.

How will the data be collected or created?

Interview data will be created firsthand by PI/with PI present. This will iavolved an interview consent form.
Construction site data will collected via Costain, the PhD co-sponsor. An NDA is in place to protect the confidentiallity of the data.

Documentation and Metadata

What documentation and metadata will accompany the data?

Interview metadata, included in already published papers, includes interviewee demographic.
Metadata of safety and quality reports will be created during the analysis and will include statistical information on data. The actual data is confidential and therefore will not be available.

Ethics and Legal Compliance

How will you manage any ethical issues?

There are minimal ethical concerns arising from this project.
Consent forms will be provided to interviewees.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

An agreement is in place.

Storage and Backup

How will the data be stored and backed up during the research?

Data will be stored on the Office 365 University system. This creates back ups of the data onto the PIs university computer and also a password protected personal harddrive.

How will you manage access and security?

Via the Office 365 system.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Both data types from this project - interview and project data - have long term value

What is the long-term preservation plan for the dataset?

They will be preserved in a University of Edinburgh data store.

Data Sharing

How will you share the data?

Raw data will not be shared due to confidentiality.
PhD thesis will contain the outputs and metadata. Effort will be made to ensure the thesis is non-confidential.

Are any restrictions on data sharing required?

Data is commercially confidential and protected by an NDA. Metadata will be provided post-analysis.

Responsibilities and Resources

Who will be responsible for data management?

PI - Henrietta Baker

What resources will you require to deliver your plan?

Data use agreement - NDA - with Costain.
Data use agreement with undergraduates participating in the research.

Appendix B

Data labelling form

Title: Safety report analysis

1. Enter unique ID.
2. Not relevant data entry.
   Indicate if data entry is not relevant to investigation.
   - Not relevant

Site attributes.
These are descriptors or attributes which can be identified BEFORE the incident occurs.

3. Which of the following OBJECTS contribute to the incident occurring?
   - Alarm
   - Alcohol / Drugs
   - Bolt
   - Cable
   - Cable tray
   - Concrete
   - Concrete liquid
   - Conduit
   - Crane
   - Door
   - Drill
   - Dunnage (light packing material eg airbag, polystrene)
   - Electricity
   - Explosives
   - Fence
   - Forklift
   - Formwork / Falsework / Shuttering
   - Gas
   - Gate
   - Grout

- Guardrail / Handrail
- Hammer
- Hand size pieces
- Hazardous substances
- Heavy material/tool
- Insect
- Hose
- Ladder
- Machinery (not self-moving)
- MEWP ('Manlift')
- Mobile phone
- Mud
- Nail
- Object on floor
- Piping
- Pontoon
- Pressure system
- Pump
- Reinforcement (bar)
- Reinforcement (mesh)
- Saw
- Site office/welfare unit ('Job trailer')
- Stairs
- Scaffold
- Screw
- Sharp edge
- Slag
- Small particle
- Spark
- Splinter / sliver
- Steel sections
- Spool
- Stud walls
- Tank eg of fuels, storage
- Timber ('Lumber')
- Tool (Powered)
- Tool (Unpowered)
- Wire
- Valve
- Vehicle (Light ie car, van)
- Vehicle (Heavy ie HGV)
- Vehicle (Plant)
- Vehicle (not specified)

- Vegetation (bushes, trees etc)
- Unpowered transporter (eg wheelbarrow)
- Water
- Wrench
- Other (free text box)

4. Which of the following WORK SITE DESCRIPTORS contribute to the incident occurring?
    - Adverse low temperature
    - Car park
    - Confined work space
    - Congested work space
    - Excavation
    - Flooding
    - Heat source
    - Heavy rain
    - Insufficient edge protection
    - Isolated
    - Lightning
    - Object at height
    - Poor housekeeping
    - Road
    - Slippery surface
    - Poor visibility
    - Uneven surface
    - Unstable support surface
    - Wind
    - Working below elevated work space
    - Working overhead
    - Working at height
    - Other (free text box)
5. Which of the following PERSONNEL DESCRIPTORS contribute to the incident occurring?
    - Drunk / under the influence (tested positive)
    - Fatigued / dizzy
    - Inattention
    - Improper body position
    - Improper procedure (by personnel ’on the day’)
    - Improper security of materials
    - Improper security of tools
    - Improper PPE
    - Insufficient PPE
    - Other (free text box)
6. Which of the following ACTIONS contribute to / occur while the incident occurring?
    - Chipping

- Cleaning
- Cutting
- Driving (Cars)
- Driving (HGV, plant)
- Driving (vehicle not specified)
- Exiting / Entering
- Excavating / breaking ground (mechanical)
- Grinding
- Lifting / pulling / manipulating (manual)
- Lifting (by machinery)
- Lifting/moving material (not specified if mechanical or manual)
- Manual digging / excavating
- Repetitive motion
- Striking / Stripping (i.e. formwork / shuttering)
- Walking / moving around
- Welding
- Other (free text box)

Consequence.
These are descriptors of the incident consequence.

7. Incident ’immediate cause’
   Need a good reason to disagree with those on site.
   - Agree with ’incident sub category’ on original data
   - Anti-social behaviour
   - Theft / vandalism
   - Unauthorised/denied entry to site
   - Intentional injury by person (violence, fights)
   - Unintentional or intent unknown injury by person
   - Verbal abuse
   - Injured by an animal
   - Insect bite/sting
   - Contact with moving plant / vehicle
   - Road Traffic Collision
   - Collapse of building or structure / structural
   - Collapse of scaffolding
   - COSHH - Escape of substances
   - COSHH - Exposure to or contact with
   - COSHH - Escape of flamable substance
   - Electric shock / Contact with electricity
   - Explosion or fire
   - Fall from height
   - Falling dust / debris into eyes

- False / unnecessary alarm
- Handling, lifting, carrying
- Pre-existing health issues
- Health issues following exposure
- Hit/Struck by moving or falling object
- Hit/Struck by fixed or stationary object
- Housekeeping (bad)
- Lifting machinery (collapse of)
- Lifting machinery (failure of load)
- Lifting machinery (overturning)
- Materials falling
- Non work related injury (e.g. taken sick at work)
- Procedural difficiencies
- Slip / Trip / Fall on same level
- Stress / anxiety / panic attack
- Trapped by something collapsing
- Unknown services - (Please specify in other box - Gas, Telecoms, Water Clean, Water dirty, Electric)
- Known services disturbed - (Please specify in other box - Gas, Telecoms, Water Clean, Water dirty, Electric)
- Unsafe conditions
- Unsafe Plant / Equipment
- Unsafe practice
- Welfare - 'The Complaint Line'
- Other (free text box)

8. Injury Type
    - Amputation
    - Asphiyxiation
    - Bite / sting
    - Bone fracture
    - Bruising
    - Burns (Minor)
    - Burns (Serious)
    - Concussion / internal head injury
    - Contusions (graze)
    - Drowned
    - Internal injury (not head)
    - Dislocation without fracture
    - Electric shock
    - Foreign object in eye
    - Injury not specified (RIDDOR 'Other not known')
    - Lacerations and open wounds
    - Loss of consciousness

- Loss of sight (temporary)
- Loss of sight (permanent blinding)
- Natural causes (e.g. common cold, flu, headache)
- Pre-existing health condition (e.g. heart failure)
- Scalping
- Skin problems (e.g. rash, allergies)
- Strains and sprains
- Superficial injuries
- ((Crush - eliminated as it is a combo of internal / fracture / bruising))
- Other (free text box)

9. Body part
    - Arms - Upper extremity
    - Back
    - Eye
    - Face
    - Finger
    - Foot
    - Legs - Lower extremity
    - Hand
    - Head
    - Torso
    - Knee
    - Shoulder
    - Wrist
    - Ankle
    - Elbow
    - Other (free text box)

Appendix C

Report for TWI

Automated Keyword Prediction using NLP (Natural Language Processing)

Henrietta Baker

Email: henrietta.baker@ed.ac.uk

November 2019

THE UNIVERSITY *of* EDINBURGH

CONTENTS

EXECUTIVE SUMMARY

The project outlined in this report was a co-operative project between Henrietta Baker, a PhD candidate at the University of Edinburgh and the Information Services team at TWI (The Welding Institute). This project was instigated by the Information Services team at TWI who were looking for novel methods to automate some of their internal processes. One of these processes was to automatically add keywords from their list of keywords to the abstracts in their report database. TWI provided the data, which consists of report abstracts with their associated keywords, while Ms Baker developed the method and code to analyse and predict the keyword classifications.

This report provides an overview of the main theory and literature in the area of using Natural Language Processing (NLP) for automatic free-text classification before outlining the methods used in this project.

Natural Language Processing (NLP) defines the suite of methods and techniques to convert human language (written or spoken) into structured data forms which can be analysed and 'understood' by computers. In this investigation, two methods of converting text are investigated: 'Bag of Words' (BOW), a statistical method involving counting words present in a piece of text; and word embedding, a deep learning method which allows greater semantic capture and preserves word order.

The results given show that for keywords which have high positive counts, i.e. the keyword occurs many times in the existing database, Gated Recurrent Unit (GRU) deep learning method prove best; however, this method is ineffectual at low positive count examples. While many deep learning methods require GPU processing power, both a fast running GPU-processed GRU algorithms and slower running CPU version are used here. This is to allow use of deep learning prior to TWI library staff arranging access the GRU hardware.

It is recommended, for Bag of Word with Machine Learning (BoW + ML) keyword prediction, to use Gradient Boost and SVM classification algorithms separately to give TWI staff flexibility and maximise the F1 accuracy for operational runtimes. It is also recommended that, if investigated further, ensemble algorithms are considered further, such as stacking algorithms.

Also, included in this report are how-to guides for downloading Python via the Spyder IDE and a manual for use of the keyword classification process at TWI for new abstracts.

INTRODUCTION & MOTIVATION

Introduction

The project outlined in this report was a co-operative project between Henrietta Baker, a PhD candidate at the University of Edinburgh and the Information Services team at TWI (The Welding Institute). TWI provided the data, which consists of report abstracts with their associated keywords, while Ms Baker developed the method and code to analyse and predict the keyword classifications.

This report begins by providing an overview of the main theory and literature in the area of using Natural Language Processing (NLP) for automatic free-text classification before outlining the methods used in this project. The results given include Bag-of-Words and word vector text representations. Recommendations for implementation in the Information Services team at TWI are provided. Also included is a how-to guide for downloading Python via the Spyder IDE.

Motivation

This project was instigated by the Information Services team at TWI who were looking for novel methods to automate some of their internal processes. One of these processes was to add keywords from their list of keywords to the abstracts in their report database.

Aim

The aim of this piece of work was to explore the different classification task methods using natural language processing (NLP) to automate keyword selection for TWI abstracts. It should be noted that it was not the aim of the project to produce a separate software to automate the process, rather to identify the best method and produce a computer code which could be used by TWI staff.

BACKGROUND AND THEORY

This section outlines Natural Language Processing (NLP) theory and defines the machine learning methods used for classification.

Much of these descriptions are taken from the following paper available on arXiv (currently in review for Automation in Construction): Henrietta Baker, Matthew R. Hallowell and Antoine J.-P. Tixier, 'Learning Construction Injury Precursors from Text', http://arxiv.org/abs/1907.11769. Specific information about the mathematics behind these algorithms can be found in the User Guide documentation for their implementation in Python's machine learning toolkit 'scikit', found at: https://scikit-learn.org/stable/user_guide.html, and links from Python's deep learning module 'keras', found at: https://keras.io/.

Classification tasks with NLP

Natural Language Processing (NLP) defines the suite of methods and techniques to convert human language (written or spoken) into structured data forms which can be analysed and 'understood' by computers. Development in text classification tasks has taken off in the last decade as the volume of written text available digitally has grown exponentially and people worry about 'data overload'.

In the context of NLP for classification of documents, generally, NLP methods are used to transform the unstructured text data into a structured format which can be passed to classification processes using machine learning prediction methods.

Early text representations relied on hand-written lexical rules which computers could follow to transform the text. This was found in most cases to be unwieldy due to word ambiguity and grammatical complexity, giving rise to the popularity of empirical language models in the late 1980s. Since then, such empirical models, based on the Bag-of-Words (BoW) representation (also known as the vector space representation), have occupied the limelight owing to the notable results found when trained on large quantities of data. Recently, another way of representing text has also emerged. Known as 'word embedding', these representations rely on dense word vectors which can then be concatenated in turn to represent a piece of text. The following subsections describe these two text representation methods.

Text Representation as Bag-of-Words Vector

The most commonly used method to create a vector representation of a written text is the 'bag-of-words' (BoW) method. Translation from free-text to a bag-of-words vector is a purely statistical method and involves counting the words present then inputting the counts into a vector V long where V is the vocabulary size and each column is a unique token in the training set, typically words but may also include punctuation, numbers etc. An extremely sparse vector is produced.

Often an aspect of pre-processing is undertaken prior to translation. This could include lowercasing all words, removing certain characters or elements (e.g. punctuation, numbers), and lemmatization or stemming (i.e. reducing a word to its base form, for example, walk, walked, and walking all have the base lemma 'walk').

A limitation of this BoW process is that semantically similar words (e.g. 'chair' and 'seat') occupy separate, orthogonal dimensions in the vector space and therefore no semantic similarity is reflected in the vector representation. Additionally, the resultant bag-of-words does not capture word order. These limitations restrict the semantic meaning which can be gained from such representations. For example, 'the weld had signs of fatigue' has the same bag-of-words representation as 'the fatigue of signs had weld'.

There are several methods which can be used to negate the information loss in this transformation. To capture word order locally, combinations of tokens, known as n-grams, may be used instead of single tokens. But doing so makes the vector space become so large and sparse that it makes it hard to fit any model, a problem colloquially known as the curse of dimensionality. In practice, it is rarely possible to use n-grams with n > 3. N-grams were not used in this analysis. Also, some syntactic information may be captured by creating different dimensions for the different Parts-Of-Speech tags of a given unigram (noun, adjective, verb, etc.), but this has the same adverse effects on the dimensionality of the space as that previously mentioned.

Text Representation as Deep Learning Vector aka 'word embedding'

Another method which can be employed to address the information loss associated with BoW is 'word embedding'. Each word is instead represented as a vector itself where the dimensions represent features shared by all words. These vectors are known as word embeddings and, unlike the long 'bag-of-words' vectors, are short (typically 100-500 dimensions) and dense. After training (supervised or unsupervised), these word embeddings allow similar words to be represented by similar vectors, encoding semantic and syntactic similarities. Several open source embedding vector databases exist, such as GloVe[1] trained on Wikipedia and Twitter entries, as well as code implementations to embed words from new text sources, such as Word2Vec[2] developed by Google. Pre-processing is less intensive than for bag-of-words methods as it is not necessary to remove stop words or stem the words; although some data cleaning may be required (spelling checks, adding spaces etc).

These word vectors can then produce a vector to represent the entire document using deep learning methods. An input piece of text is then represented as a sequence of word vectors. Deep Learning architectures take these vectors as input and pass them through their layers. Each layer computes a higher-level, more abstract representation of the input text by performing operations (e.g. convolutional, recurrent) on top of the output of the previous layer, until a single vector representing the entire input is obtained. This task is known as representation learning.

The resultant dense vector not only represents the words which occur but also their order and the sematic relationships. This contrasts with the long, sparse vector produced in bag-of-words representations which only captures word counts. The word embedding vector can be passed to the machine learning methods in the same way the BoW sparse vector could, or a final layer (such as a softmax layer) included in the deep learning method used for representation learning to classify the text.

Machine Learning Classification Methods - Supervised Learning

Machine learning, a term first introduced by Arthur Samuel in 1959, is now used to describe a wide range of computer-based learning tasks which employ mathematical algorithms and statistical models to carry out tasks. These computational models rely on patterns and inference, rather than explicit instructions, to achieve their aims.

Predicting keywords from abstract text can be viewed as a classification task where each keyword is a class. Classification tasks can have binary or multiple classes. Binary classes are characterised by only two possible class options, i.e. document is either class A or class B. Multiple classes, where more than two classes exist, are more computationally intensive and require different methods.

Considering the task here, it can either be viewed as a multi-class task where each keyword is a separate class or as series of binary classification tasks where a document either belongs to the class containing the keyword or belongs to the class without that keyword. In the second case, each keyword is an independent, separate classification task. The analysis presented here undertakes this binary classification as there are too many keywords to successfully undertake multi-class analysis.

All methods undertaken in this analysis are supervised classification tasks, meaning that they use training data, where the outcome/classification is known, to train the parameters of the model. The performance of these models is then tested on a test set which has been kept separate from the training data. The final reported accuracy values are from validation data which was kept separate throughout.

[1] Found at: https://nlp.stanford.edu/projects/glove/
[2] Found at: https://code.google.com/archive/p/word2vec/

Naïve Bayes

This supervised method uses the BoW text representation and applies binary keyword classification to work out probability of keyword assigned, or keyword not assigned. The method naively (hence the name) assumes independence between all inputs (words in this case) and applies Bayes Rule which states:

$$P(A|B) = \frac{P(A)\times P(B|A)}{P(B)} \quad \text{i.e. } Probability\ of\ A\ given\ that\ B\ is\ present\ = \frac{Probability\ of\ A \times Probability\ of\ B\ given\ that\ A\ is\ present}{Probaility\ of\ B}$$

The probabilities on the right-hand side of this equation are calculated from the observed frequencies in the training data. In this case:

- P(A) = the probability of the keyword.
- P(B) = the probability that the abstract text is composed.
- P(B|A) = the probability of the abstract text given that the keyword is assigned.
- P(A|B) = the probability of the keyword given the abstract text – this is what we want to calculate in new cases.

For example, if we have a training set of 100 and 30 of them are assigned keyword A then P(A) = 30/100=0.3. For P(B), this requires more computation as no exact text will be seen twice. Therefore, the word frequencies are used such that P('this is a sentence') = P('this') * P('is') * P ('a') * P('sentence').

A worked example is included for clarification. In this example, the observed frequencies are recorded in Table 1 (e.g. in the training set, Keyword A is observed 30 times and 'weld' is seen 22 times, twice in the set assigned Keyword A and 20 in the set not). These frequencies are used to calculate the probabilities.

Table 1: Observed Frequencies in Training Set of Naïve Bayes Worked Example

Number of training examples **n = 100**	**Keyword A**	**Not Keyword A**
	30	70
Iron	5	15
Weld	2	20
Rust	10	5

Given this simple example, for an abstract containing only the words 'iron', 'weld' and 'rust':

$$P(A) = \frac{30}{100} = 0.3 \ \text{ and } \ P(\bar{A}) = \frac{70}{100} = 0.7$$

$$P(B) = P('iron\ weld\ rust') = P('iron') \times P('weld') \times P('rust') = \frac{20}{100} \times \frac{22}{100} \times \frac{15}{100} = \frac{66}{10000} = 0.0066$$

$$P(B|A) = P('iron\ weld\ rust'|A) = \frac{5}{30} \times \frac{2}{30} \times \frac{10}{30} = \frac{1}{270} = 0.0037$$

$$P(B|\bar{A}) = P('iron\ weld\ rust'|\bar{A}) = \frac{15}{70} \times \frac{20}{70} \times \frac{5}{70} = \frac{15}{3430} = 0.00437$$

Therefore, $P(A|B) = \frac{P(A)\times P(B|A)}{P(B)} = 0.168$ while $P(\bar{A}|B) = \frac{P(\bar{A})\times P(B|\bar{A})}{P(B)} = 0.463$

As $P(A|B) < P(\bar{A}|B)$ then the model would predict that the abstract is not assigned keyword A.

In this way, probabilities are calculated using the training data observed frequencies.

An issue with this method occurs when a word appears in the test text which wasn't observed in the training data. If $P(word) = 0$, then $P(text) = P(B) = 0$, $P(B|A) = 0$ as well as $P(B|\bar{A}) = 0$. Smoothing techniques can be employed to overcome this. The most common, known as 'additive smoothing', simply adds an observed count to every word in the final vocabulary.

k-Nearest Neighbour (kNN)

k-Nearest Neighbour algorithm operates on the assumption that the vector representation of the text will be located close to others of the same output (keyword).

The cosine distance between the new text vector and existing vectors is calculated at the prediction stage and the nearest point is found. The keywords are then provided by the nearest point found. This method is computationally low during learning as it requires no machine learning phase; however, this means that the prediction phase is high computationally for large existing examples.

This method does not need smoothing (unlike Naïve Bayes); however, the comparative document length impacts the analysis. Therefore truncating/extending/scaling the vector number is commonplace so that all the text inputs (therefore vector representations) have the same number of features.

In the simple example given in **Error! Reference source not found.**, the new point is not assigned Keyword A as its represented vector is closest to a 'Not Keyword A' existing point. This is using 1NN, aka next-nearest neighbour, as it only takes the nearest point into account. Other variations of the algorithm use techniques like weighted averages to avoid basing decisions on outliers etc.

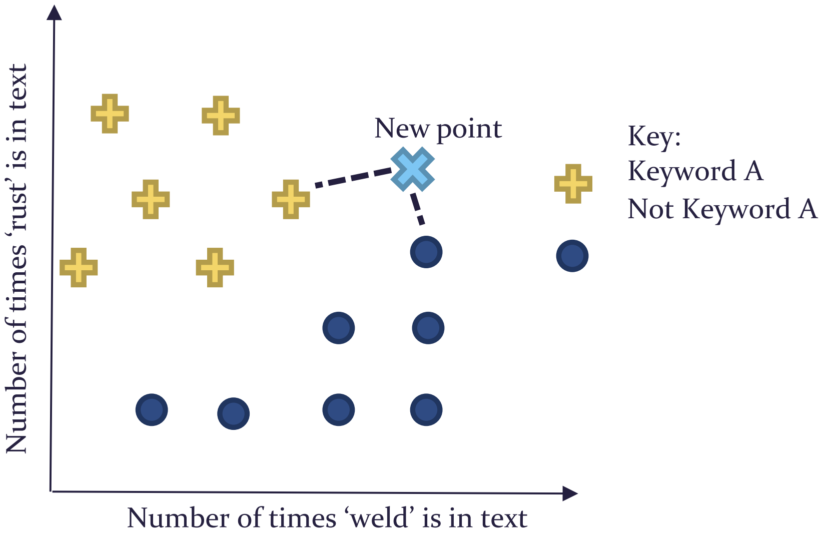


Figure 1: Simple kNN Example

Decision Tree

Decision tree method for classification produces a tree of binary decisions which lead to the classification. They are also known as CART (classification and regression tree) methods. Each level of a simple decision tree asks a binary (yes/no) question which splits the data depending on its features. Each branch question is selected by the greatest entropy change i.e. what can split the data the most. This method is often used in high accountability industries, such as medicine and finance, as it is easily interpretable and to explain why a certain decision was made. This method can be used for multiple class classification and regression as well as binary classification.

An example of this is given in **Error! Reference source not found.**.

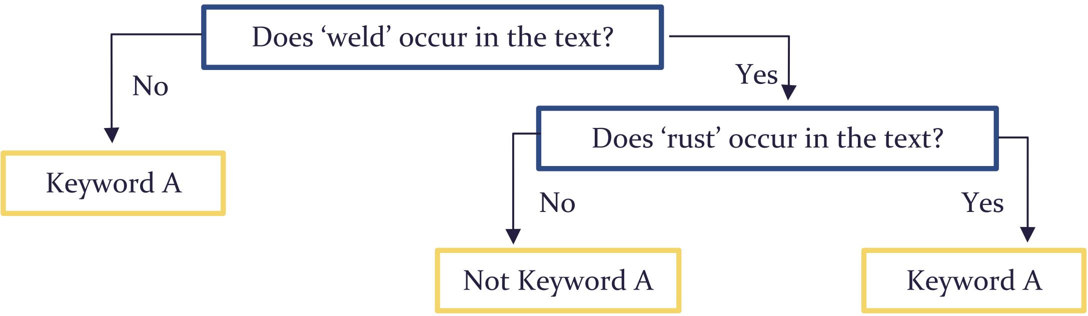

Does 'weld' occur in the text?
No
Yes
Keyword A
Does 'rust' occur in the text?
No
Yes
Not Keyword A
Keyword A

Figure 2: Decision Tree Example

Gradient Boosting

Gradient boosting is an ensemble technique that aims to reduce error in decision tree methods by adding shallow decision trees (weak high bias-low variance base models) in sequence, repeatedly reducing the bias of the entire sequence.

Python *sklearn* module's implementation of gradient boosting classifier was used in this analysis.

[NB. Gradient boosting contrasts with Random Forest algorithms (another ensemble algorithm for decision trees) which aim to reduce error by decreasing the variance by building deep decision trees (complex low bias-high variance base models) in parallel. Random Forest ensemble was not used in this investigation.]

Linear Support Vector Machine (SVM)

Arguably the most popular classification algorithm, linear SVM is a geometrical method that seeks to classify points into two categories by finding the best separating hyperplane. 'Best separating' can be defined in many ways; however, it essentially aims to draw a line between the data with the largest gap either side to the data categories. New data is then classified by which side of the hyperplane its representative vector lies.

In a simple 2D example, as seen in Figure 3, this hyperplane is a line and, in the example, the new data point is assigned Keyword A. It should be noted that this is the opposite prediction to that given in the kNN example. This shows how two methods, given the same data, can make different decisions.

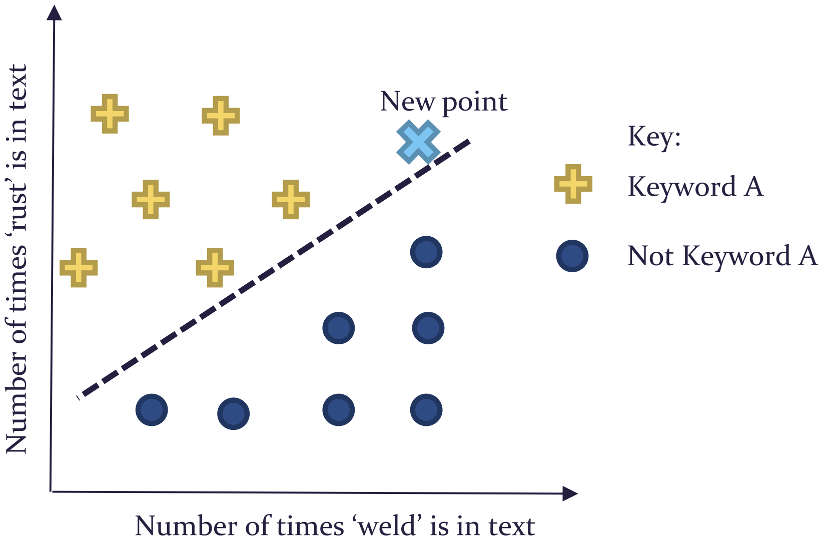

Number of times 'rust' is in text
New point
Key:
Keyword A
Not Keyword A
Number of times 'weld' is in text

Figure 3: Linear Support Vector Example

Bagging aka Bootstrap Aggregating

Bagging, or bootstrap aggregating, is an ensemble method which aims to increase accuracy by decreasing the variance in the model. It also mitigates against overfitting. While generally used on CART (classification and regression trees), it can be used on any model and is a type of model averaging.

Bootstrap aggregating involved taking bootstrap samples of the data to train multiple predictors and then averaging the trained variables to create a single, more stable predictor. A bootstrap sample is generated by random selection with replacement.

[NB: 'Random Forest' ensemble method for CART models (mentioned earlier) uses principles of bagging.]

Deep Learning Method

Deep learning refers to machine learning methods with one or more hidden layer of variables. These can include some deep generative models; however, they most commonly refer to neural networks.

Neural networks were originally inspired by the biological processes which neurons exhibit in brains. A neural network is characterised by a defined number of layers linking input to output where at least one of these layers is 'hidden'. A classic neural network model can be represented as in Figure 4, where the input features are linked to the output through a series of nodes. The deep learning process then uses labelled training data to train the variables (aka weights) along each link and at each node. This is a very simple overview and more information can be found in any deep learning textbook.

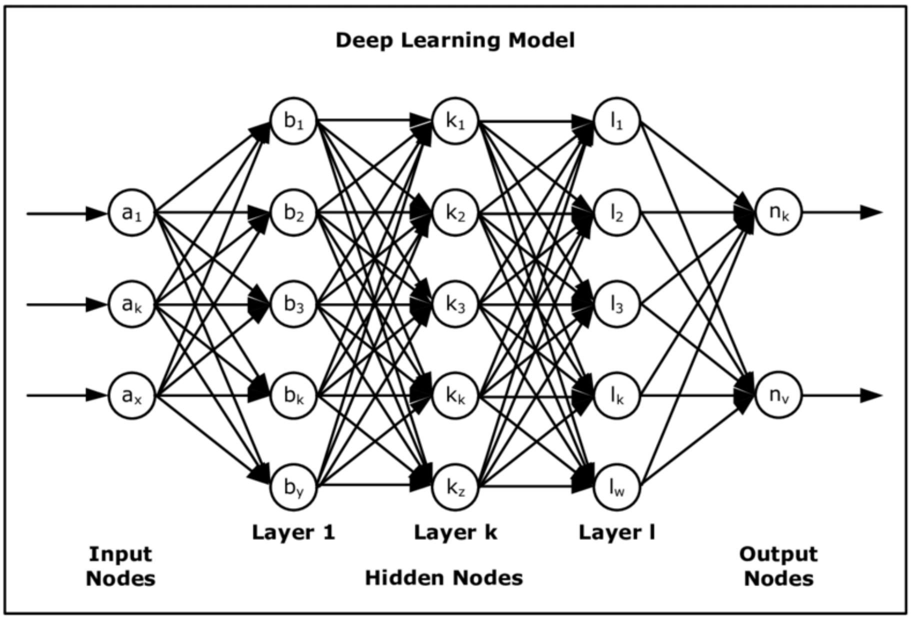


Figure 4: Neural Network Model © Will Serano, Smart Internet Search with Random Neural Networks, European Review 25(02):1-13, February 2017.

However, this type of neural network (also known as a convolutional neural network) does not preserve the feature order which is extremely important for natural language processing (NLP). Therefore, this investigation uses a type of Recurrent Neural Network called Gated Recurrent Unit (GRU), introduced by Cho et al in 2014 [3].

Recurrent Neural Networks (RNNs) have a loop in them such that information from a previous input affects the next and so on. This makes these methods a good choice for dealing with natural language inputs. GRU is a variant of the LSTM (long short-term memory) method but is simpler and faster than it's forebearer. Information on the mathematics behind these methods can be found linked from documentation for Python's deep learning module 'keras', found at: https://keras.io/ and in the paper in footnote 3 below.

[3] Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". arXiv:1406.1078

Accuracy metrics

Three accuracy metrics are generally used to assess the performance of a classification task: precision, recall and F1, the harmonised average of the previous two.

Precision measures whether a label predicted by the model is correct or not.

Recall measures the proportion of correct labels assigned.

F1 is a harmonised average of the two.

An example to illustrate this is as follows:

Of 100 abstracts, 30 have keyword 'steel' assigned. The model predicts 35 abstracts assigned 'steel'; 15 of which are incorrect. These values can be represented in Table 2.

Table 2: Accuracy metrics example

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive 20	False Negative 10
	Negative	False positive 15	True negative 55

Precision, the measure of how correct the assigned labels are, is defined as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} = \frac{20}{35} = 0.571$$

Recall, the measure of how many of the labels were predicted, is defined as:

$$Recall = = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{20}{30} = 0.667$$

F1, the harmonic average of the two, is defined as:

$$F1 = 2 \times \frac{Precision\ \times\ Recall}{Precision + Recall} = 2 \times \frac{0.571 \times 0.667}{0.571 + 0.667} = 0.615$$

The harmonic average is used as it penalises extremely low values of recall and precision, and so is a good measure of the operational usefulness of the model. This is especially relevant when binary class data is extremely unbalanced, i.e. where one class is much bigger than the other. For example, in a dataset where one class is only 1% of the data, accuracy values of 99% can be obtained by simply never assigning that class. F1 penalises this and forces the model to go to a middle ground.

Further approaches/information

Computing requirements

Deep learning approaches often require the use of GPU processing – NVIDIA GTX 1050 was used for this analysis. A link to instructions to set up GPU processing for python is included in the appendix. Setting up a subscription account with a cloud computing provider (for example, Microsoft's Azure service) is another option for accessing computing resource without purchasing in-house hardware.

A deep learning option not requiring GPU processing is included in this investigation.

Unsupervised ML classification methods

Unsupervised ML classification methods were considered unsuitable for this task due to the high fidelity desired by the process and the readily available labelled dataset required for supervised machine learning methods.

METHODS

This section outlines the methodology and methods employed in this piece of work.

Methodology

The methodology adopted for this piece of work is positivist; however, there are three assumptions which should be highlighted: (1) that the keyword labels are consistent throughout the dataset; (2) that the keywords are all present and correct; (3) keywords are independent.

(1) 'keyword labels are consistent throughout dataset'
The methods adopted in this work assume that keywords are assigned consistently throughout the data. However, in discussion with TWI staff, it came to light that over time more keywords have been added to the collection without post-labelling previously added articles. This would result in 'false negatives' within the data. False negatives in the training data could affect the accuracy of training the model, while false negatives in the test data will affect the accuracy assessment.

(2) 'keywords are all present and correct'
No allowance was made for incorrect keyword assignation in the input data. Similar to above, this can affect the quality of the trained model and accuracy assessment criteria.

(3) 'keywords are independent'
As previously mentioned, the task is set as a series of binary classification tasks. This assumes keyword independence. This simplifies the algorithms required, especially due to the number of categories.

To ensure validity of the results, a select set was randomly set aside at the beginning of the project as the validation set (15% of total abstracts). The remaining 85% were randomized and divided using 20-80 validation, where the model is trained on 80% of the data and tested on 20%. All accuracy results reported on in this report are validation accuracies.

'Bag-of-Words' approach

Initially BoW vector method was used and the following machine learning methods were employed: Naïve Bayes, kNN, Decision Tree, Gradient Boosting and SVM.

Transformation to BoW vector

To process the data into a Bag-of-Words representation, the following step were taken:

1. 0.6% of the dataset does not contain an abstract; they either they included the text: 'No abstract', 'No abstract available, or no text at all. Therefore, during pre-processing, the short abstracts (n<5) were removed.
2. The abstract text was extracted and 'tokenised' - meaning split into individual words by white space.
3. All words were lowercased, and punctuation/numbers were removed.
4. The words were stemmed, using the SnowballStemmer, which takes words back to their root stem. E.g. 'welding' to 'weld'.
5. Stopwords were removed. Stopword list from English stoplist in Python's *nltk* (Natural Language Toolkit) module.
6. 'Extreme' word counts removed. Only words which occur more than 5 times or less than 10000 times in the entire data set were included in the BoW vector. This is found to be ~18500 words.
7. TF-IDF (term frequency–inverse document frequency) weightings were applied to the words.
8. Each abstract vector was normalised to account for their variation in length. NB: this means abstract length is removed as a feature.

Algorithms and Optimisation

Listed below are the Python's scikit-learn implementations of these algorithms used:

- Naïve Bayes - sklearn.naive_bayes.MultinomialNB
- kNN - sklearn.neighbors.NeighborsClassifier
- Decision Tree - sklearn.tree.DecisionTreeClassifier
- Gradient Boosting - sklearn.ensemble.GradientBoostingClassifier
- SVM - sklearn.linear_model.SGDClassifier
- SVM Bagging - sklearn.ensemble.BaggingClassifier

These algorithms were implemented and roughly tuned for overall optimisation using a sample of the keywords. These algorithms were coarsely tuned for overall optimisation using a sample of 150 keywords. This means that each algorithm was optimised for the sample keyword set, not for each individual keyword. These settings are implemented in the classification file. Individual optimisation is recommended for further investigation.

Random Forest algorithm was not included in this investigation as initial trials found the accuracy benefits were minimal and the runtime was too long to be operational. However, further investigation using an optimised version of random forest may prove useful.

Deep Learning approach

Deep learning using neural network This investigation first used a fast GRU (Gated Recurrent Unit) keras implementation using TensorFlow backend.[4] A later edit changed this to the ordinary implementation of GRU keras so that the prediction would run on CPU. The full documentation for keras can be found at: **https://keras.io/**.

The two implementations used are:

- GRU (Fast) - keras.layers.CuDNNGRU
- GRU - keras.layers.GRU

The following steps are taken:

1. Abstract text was lower-cased, and punctuation removed.
2. Abstract text was pre-processed using keras' text preprocessing script and word sequences padded/truncated to 100 words. [NB: This length is an input and can be edited in the code]
3. An embedding layer is then implemented to convert the raw text into a dense input matrix.
4. A bidirectional GRU is implemented. While originally a fast implementation of this algorithm was used, this requires GPU processing power which was not easily accessible to the TWI team. Therefore, a slower running (but equally accurate) GRU algorithm was used.

[4] Script based on https://www.kaggle.com/sudalairajkumar/a-look-at-different-embeddings/data

RESULTS

Data Exploration

Initial exploration of the data distribution found that of the 1945 unique keywords used to label the abstracts, 913 – nearly half all keywords - occurred in less than 0.1% of the abstracts. Only 183 keywords occurred in more than 2% of abstracts. The full distribution can be seen in Figure 5.

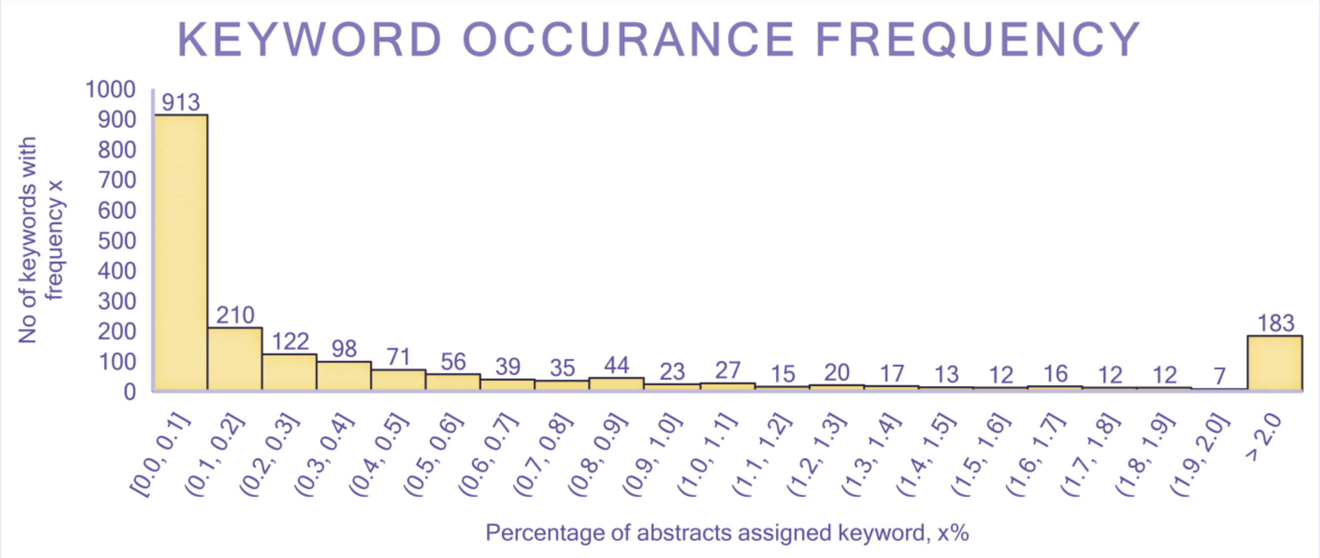


Figure 5: Graph showing the distribution of keyword frequencies

These sparse positive examples of keywords pose an issue for machine learning methods. This is known as an **imbalanced classification problem**. In fact, if a class occurs less in than 5% of the total dataset, this is normally referred to as a rare class and ML algorithms struggle to achieve high accuracy values. As 96% of the keywords fell in this regime, imbalanced class frequency in this dataset proved to be a hinderance to achieving high accuracies, especially recall values.

There are two main collections of methods used to mitigate the effects of imbalanced dataset and improve the accuracy: data resampling or algorithm improvements.

Resampling essentially aims to balance the dataset more evenly by either increasing the minority examples used in training or decreasing the majority one. There are several ways in which this can be achieved, for example, oversampling the minority class or undersampling the majority class. In this investigation, the minority class (those abstracts assigned the keyword) were oversampled to balance of the dataset.

Algorithm improvements can be implemented to specifically deal with imbalanced datasets. These generally are the form of ensemble algorithms, where many classifiers are combined to create a stronger overall classification. Gradient boosting and bootstrapping, as used in this investigation, are examples of an ensemble classifier. Other methods, such as model stacking, were briefly investigated; however further investigation may prove fruitful in increasing accuracy.

BoW Results

Figure 6 and Figure 7 show tables of the accuracy values for a sample of 150 (out of 1945) keyword classification tasks.

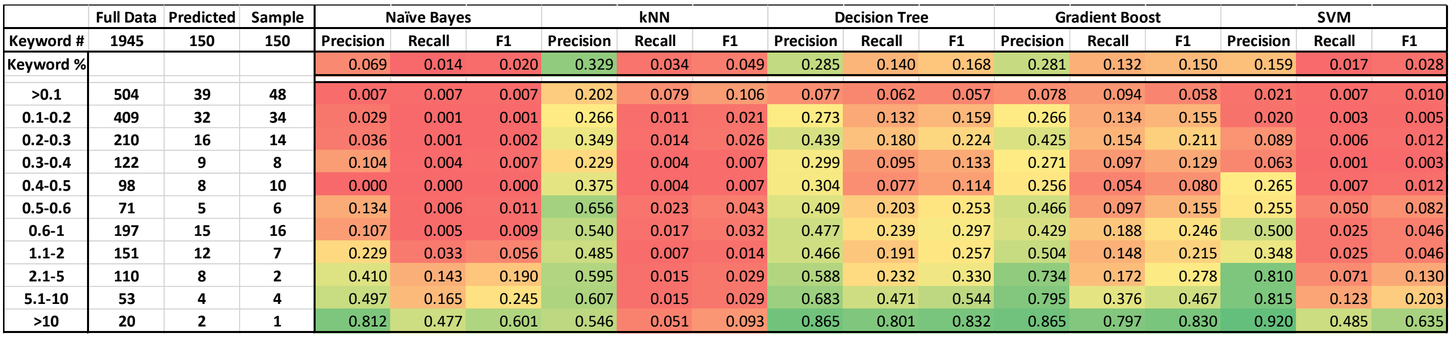
As a results of a small optimisation task on the train-test dataset, a minimum minority class proportion of 10% was selected for the data oversampling. This means that the keyword-assigned abstracts were replicated until a minimum of 10% of the training set was composed of abstracts assigned the keyword. Normally, oversampling would be done at random; however, as the dataset was extremely unbalanced, in this investigation the positive examples were replicated as a whole set until the minimum 10% was achieved. For example, 'absorption' was a keyword in 588 abstracts. This is ~0.4% of the entire training dataset. Therefore, each of these abstracts was included 25 times in the training set to bring up the proportion to ~10%. The benefits of this are clearly seen in the comparison of F1 accuracy values in Figure 6 and Figure 7. It should be noted, however, that for Decision Tree methods (i.e. Decision Tree and Gradient Boost) this oversampling method decreases the precision of the model, while significantly increasing the recall.

The results in Figure 7 show that the Gradient Boost ensemble is the highest performing algorithm for the resampled data and is recommended for use in implementation. It slightly outperforms SVM, SVM bagging and the Decision Tree models. These algorithms significantly outperform Naïve Bayes and kNN which are not considered further in this investigation.

SVM and SVM bagging algorithms are more precise than Gradient Boost at higher proportion keyword assigned. In discussion with the TWI staff, it was decided that precision was important to them, or they would at least appreciate the option of choosing precision over recall and vice versa. However, although SVM accuracy performance is slightly outperformed by a Bagging (aka Bootstrap) ensemble of the same algorithm, the run time is significantly increased (~10x greater) for this. In order to maximise the operational suitability for implementation, it is recommended that SVM is used without bagging.

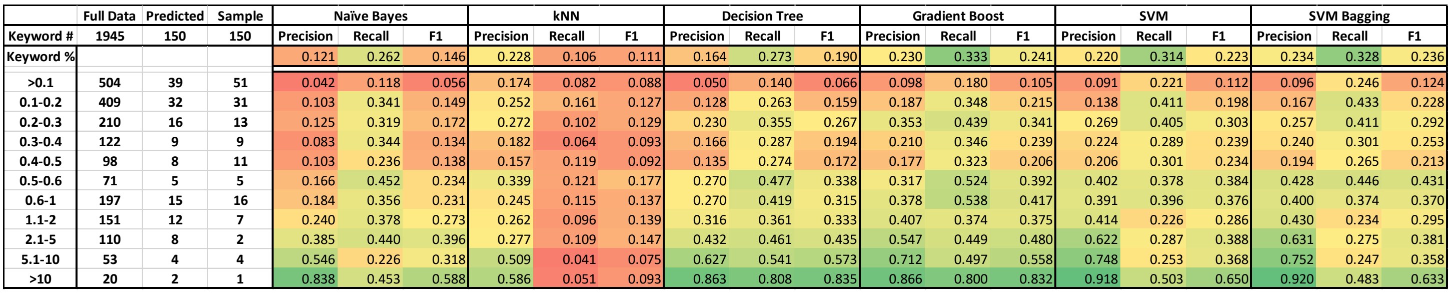
A voting algorithm, 'majority-rules', was briefly considered for investigation. This ensemble method predicts the classification using multiple algorithms and then uses a majority vote decision to assign the class. However, as only two algorithm types were suitable for inclusion due to the low accuracies of the other methods, this was not successful in increasing F1 accuracy metrics but did increase the precision of the results. These results are seen in Figure 8. While it is not recommended for implementation at this point, further work could investigate further stacking ensemble methods.

To summarise, for BoW + ML keyword prediction, it is recommended to use Gradient Boost and SVM algorithms separately to give TWI flexibility and maximise the F1 for operational runtimes. It is also recommended that, if investigated further, other ensemble algorithms are considered.



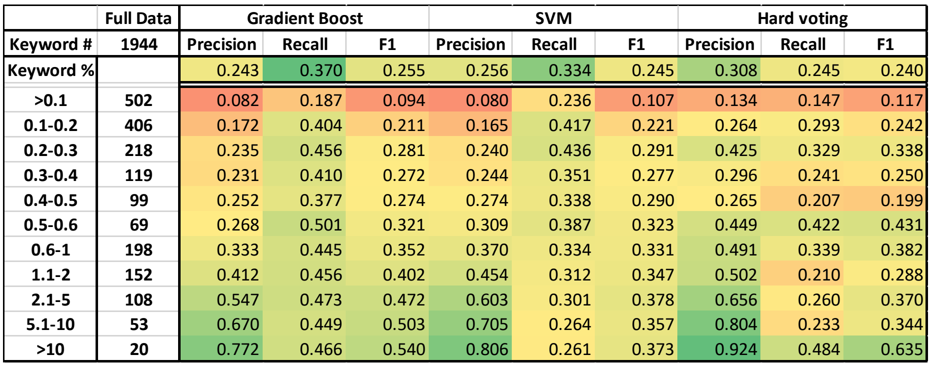
	Full Data	Predicted	Sample	Naïve Bayes			kNN			Decision Tree			Gradient Boost			SVM		
Keyword #	1945	150	150	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Keyword %				0.069	0.014	0.020	0.329	0.034	0.049	0.285	0.140	0.168	0.281	0.132	0.150	0.159	0.017	0.028
>0.1	504	39	48	0.007	0.007	0.007	0.202	0.079	0.106	0.077	0.062	0.057	0.078	0.094	0.058	0.021	0.007	0.010
0.1-0.2	409	32	34	0.029	0.001	0.001	0.266	0.011	0.021	0.273	0.132	0.159	0.266	0.134	0.155	0.020	0.003	0.005
0.2-0.3	210	16	14	0.036	0.001	0.002	0.349	0.014	0.026	0.439	0.180	0.224	0.425	0.154	0.211	0.089	0.006	0.012
0.3-0.4	122	9	8	0.104	0.004	0.007	0.229	0.004	0.007	0.299	0.095	0.133	0.271	0.097	0.129	0.063	0.001	0.003
0.4-0.5	98	8	10	0.000	0.000	0.000	0.375	0.004	0.007	0.304	0.077	0.114	0.256	0.054	0.080	0.265	0.007	0.012
0.5-0.6	71	5	6	0.134	0.006	0.011	0.656	0.023	0.043	0.409	0.203	0.253	0.466	0.097	0.155	0.255	0.050	0.082
0.6-1	197	15	16	0.107	0.005	0.009	0.540	0.017	0.032	0.477	0.239	0.297	0.429	0.188	0.246	0.500	0.025	0.046
1.1-2	151	12	7	0.229	0.033	0.056	0.485	0.007	0.014	0.466	0.191	0.257	0.504	0.148	0.215	0.348	0.025	0.046
2.1-5	110	8	2	0.410	0.143	0.190	0.595	0.015	0.029	0.588	0.232	0.330	0.734	0.172	0.278	0.810	0.071	0.130
5.1-10	53	4	4	0.497	0.165	0.245	0.607	0.015	0.029	0.683	0.471	0.544	0.795	0.376	0.467	0.815	0.123	0.203
>10	20	2	1	0.812	0.477	0.601	0.546	0.051	0.093	0.865	0.801	0.832	0.865	0.797	0.830	0.920	0.485	0.635

Figure 6: Table of accuracy for BoW + ML (No Data Scaling)



	Full Data	Predicted	Sample	Naïve Bayes			kNN			Decision Tree			Gradient Boost			SVM			SVM Bagging		
Keyword #	1945	150	150	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Keyword %				0.121	0.262	0.146	0.228	0.106	0.111	0.164	0.273	0.190	0.230	0.333	0.241	0.220	0.314	0.223	0.234	0.328	0.236
>0.1	504	39	51	0.042	0.118	0.056	0.174	0.082	0.088	0.050	0.140	0.066	0.098	0.180	0.105	0.091	0.221	0.112	0.096	0.246	0.124
0.1-0.2	409	32	31	0.103	0.341	0.149	0.252	0.161	0.127	0.128	0.263	0.159	0.187	0.348	0.215	0.138	0.411	0.198	0.167	0.433	0.228
0.2-0.3	210	16	13	0.125	0.319	0.172	0.272	0.102	0.129	0.230	0.355	0.267	0.353	0.439	0.341	0.269	0.405	0.303	0.257	0.411	0.292
0.3-0.4	122	9	9	0.083	0.344	0.134	0.182	0.064	0.093	0.166	0.287	0.194	0.210	0.346	0.239	0.224	0.289	0.239	0.240	0.301	0.253
0.4-0.5	98	8	11	0.103	0.236	0.138	0.157	0.119	0.092	0.135	0.274	0.172	0.177	0.323	0.206	0.206	0.301	0.234	0.194	0.265	0.213
0.5-0.6	71	5	5	0.166	0.452	0.234	0.339	0.121	0.177	0.270	0.477	0.338	0.317	0.524	0.392	0.402	0.378	0.384	0.428	0.446	0.431
0.6-1	197	15	16	0.184	0.356	0.231	0.245	0.115	0.137	0.270	0.419	0.315	0.378	0.538	0.417	0.391	0.396	0.376	0.400	0.374	0.370
1.1-2	151	12	7	0.240	0.378	0.273	0.262	0.096	0.139	0.316	0.361	0.333	0.407	0.374	0.375	0.414	0.226	0.286	0.430	0.234	0.295
2.1-5	110	8	2	0.385	0.440	0.396	0.277	0.109	0.147	0.432	0.461	0.435	0.547	0.449	0.480	0.622	0.287	0.388	0.631	0.275	0.381
5.1-10	53	4	4	0.546	0.226	0.318	0.509	0.041	0.075	0.627	0.541	0.573	0.712	0.497	0.558	0.748	0.253	0.368	0.752	0.247	0.358
>10	20	2	1	0.838	0.453	0.588	0.586	0.051	0.093	0.863	0.808	0.835	0.866	0.800	0.832	0.918	0.503	0.650	0.920	0.483	0.633

Figure 7: Table of accuracy values for BoW + ML (Data scaled to 10% positive examples)



	Full Data	Gradient Boost			SVM			Hard voting		
Keyword #	1944	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Keyword %		0.243	0.370	0.255	0.256	0.334	0.245	0.308	0.245	0.240
>0.1	502	0.082	0.187	0.094	0.080	0.236	0.107	0.134	0.147	0.117
0.1-0.2	406	0.172	0.404	0.211	0.165	0.417	0.221	0.264	0.293	0.242
0.2-0.3	218	0.235	0.456	0.281	0.240	0.436	0.291	0.425	0.329	0.338
0.3-0.4	119	0.231	0.410	0.272	0.244	0.351	0.277	0.296	0.241	0.250
0.4-0.5	99	0.252	0.377	0.274	0.274	0.338	0.290	0.265	0.207	0.199
0.5-0.6	69	0.268	0.501	0.321	0.309	0.387	0.323	0.449	0.422	0.431
0.6-1	198	0.333	0.445	0.352	0.370	0.334	0.331	0.491	0.339	0.382
1.1-2	152	0.412	0.456	0.402	0.454	0.312	0.347	0.502	0.210	0.288
2.1-5	108	0.547	0.473	0.472	0.603	0.301	0.378	0.656	0.260	0.370
5.1-10	53	0.670	0.449	0.503	0.705	0.264	0.357	0.804	0.233	0.344
>10	20	0.772	0.466	0.540	0.806	0.261	0.373	0.924	0.484	0.635

Figure 8: Table of accuracy values for BoW + ML (All keywords predicted, data scaled to 10% positive examples)

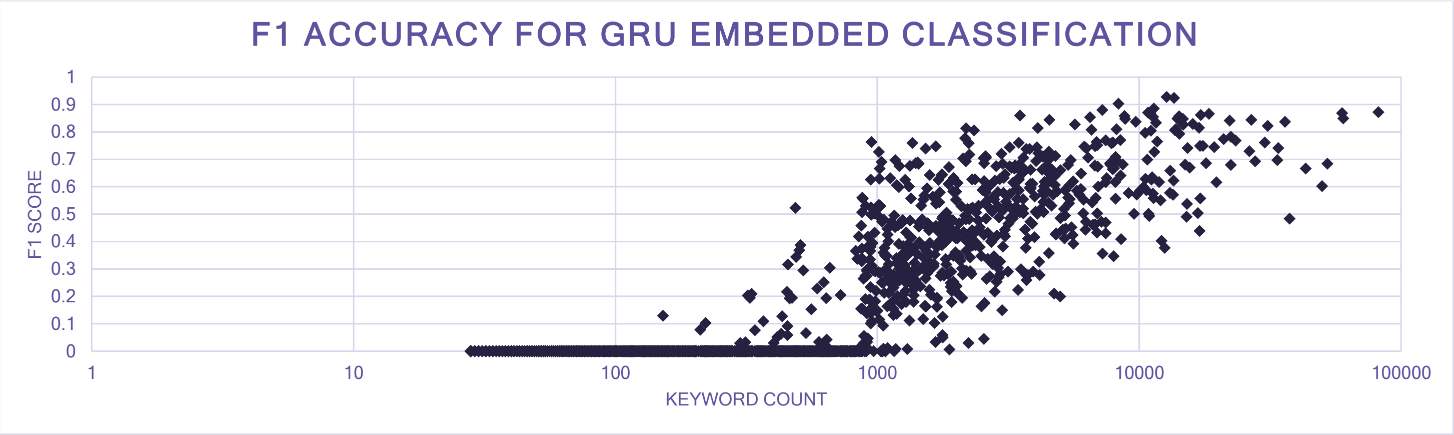


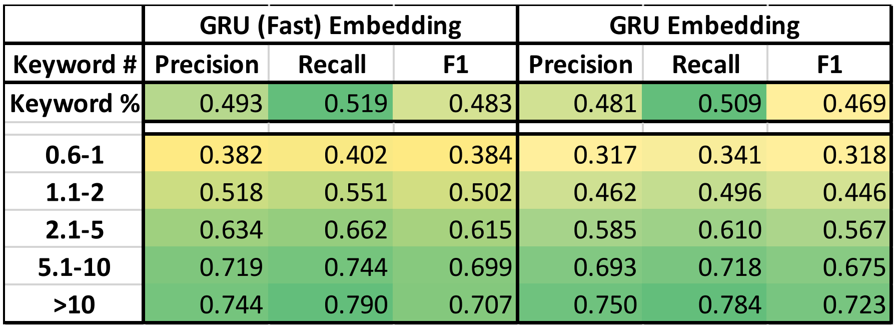
Figure 9: Graph showing F1 accuracy score for GRU classification

Deep Learning approach

Predicting keywords using the Gated Recurrent Unit (GRU) deep learning method was undertaken first using the fast GRU implementation in keras (which uses GPU processing) then repeated using a slower running GRU implementation (still in keras) so that TWI staff could start using it before they arrange access to the GPU hardware.

As seen in Figure 9, deep learning did not achieve reasonable accuracy scores for keywords with <900 positive example in the training set (~0.5% of dataset). It is therefore recommended that deep learning is only implemented on those keywords which occur >0.6% of abstracts (~1000 positive examples). The specific accuracy results for this are included in Figure 10. This demonstrates that there is a minimal accuracy loss in switching to the slower running GRU algorithm. However, it is much slower running (see timing section in User Guide Appendix).

This method outperforms the best BoW + ML method for keywords occurring in 6%-10% of the training set.



	GRU (Fast) Embedding			GRU Embedding		
Keyword #	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Keyword %	0.493	0.519	0.483	0.481	0.509	0.469
0.6-1	0.382	0.402	0.384	0.317	0.341	0.318
1.1-2	0.518	0.551	0.502	0.462	0.496	0.446
2.1-5	0.634	0.662	0.615	0.585	0.610	0.567
5.1-10	0.719	0.744	0.699	0.693	0.718	0.675
>10	0.744	0.790	0.707	0.750	0.784	0.723

Figure 10: Table showing accuracy results for fast GRU and normal implementation

CONCLUSION AND RECOMMENDATIONS

This report documented the investigation conducted into using Natural Language Processing (NLP) and machine learning (ML) methods to predict keywords from abstract text for the TWI document library. This problem was treated as a series of binary classification tasks, where each task predicted whether the abstract text belonged in the keyword class or not. This was repeated for each of the 1945 keywords present in the TWI database.

Initial investigations showed that that imbalanced class distribution would prove to be an issue during these tasks. Imbalanced class distribution is where one class contains more entries than another. In this case, for each keyword, there were far more examples where the keyword is not assigned than examples of abstracts with the keyword. Generally, in ML, if a class occurs less in than 5% of the total dataset, this is normally referred to as a rare class and ML algorithms struggle to achieve high accuracy values. As 96% of the keywords fell in this regime, imbalanced class distribution in this dataset proved to be a hinderance to achieving high accuracies, especially recall values. To mitigate this, resampling was undertaken where the abstract texts which were assigned the keyword were repeated in the training set until a minimum proportion of 10% was achieved.

Two text representation methods were used: 'Bag of Words' (BoW) and word embedding. The BoW representation was used as input into the following classification machine learning (ML) algorithms: Naïve Bayes, K-Nearest Neighbour, Decision Tree, Gradient Boosting, Support Vector Machine (SVM) and SVM Bagged (aka Bootstrap Aggregation). The scikit implementations of these algorithms were used in python. The word embedding was trained solely on the TWI abstract data (no pretrained vectors were used) and implemented a Gated Recurrent Unit (GRU) algorithm, a deep learning method, in keras.

The results showed that for keywords which occurred in 0.6%-10% of abstracts, the deep learning method gave greater average accuracies. Meanwhile, the highest performing BoW + ML algorithm was Gradient Boost. However, SVM algorithm obtained more precise prediction than Gradient Boost at higher proportion keyword assigned. Therefore, it is recommended to also include both these BoW + ML algorithms to allow give flexibility in the results for TWI staff. Prediction for keywords occurring in less than 0.6% of the database is unreliable. This represents 75% of the keywords used at TWI.

Therefore, it is recommended that keyword prediction is undertaken only for keywords which occur in >0.6% of abstracts (>1000 examples in the training data used in this investigation). The methods implemented are Gradient Boost, SVM and GRU Deep Learning. Use of multiple methods gives the TWI staff the option of comparing the output.

Not every avenue could be explored in this project. The following points are noted for possible further development:

- Keyword dependencies. This investigation assumes all keywords are independent. This is not true as keywords are hierarchical and some have significant co-occurrence. Therefore, including this consideration has potential to increase accuracy scores.
- Use of bi-grams, tri-grams etc. Including frequently observed word groups (bi-grams, tri-grams) may increase accuracies, but will add to the dimensionality of the Bag-of-Words model.
- Adding a keyword search element. Current predictions are done entirely using the text representations and prediction algorithms. It may increase the recall to include a keyword search module; such that if the keyword text is present in the abstract, it is predicted.
- Filtering for those keywords introduced later (removing false negative from training set). In talks with the TWI library staff, it came to light that some keywords were added to their possible keyword list later and were not back assigned to records already in the database. This would result in some 'false negatives' i.e. records which should be assigned the keyword but aren't, in the

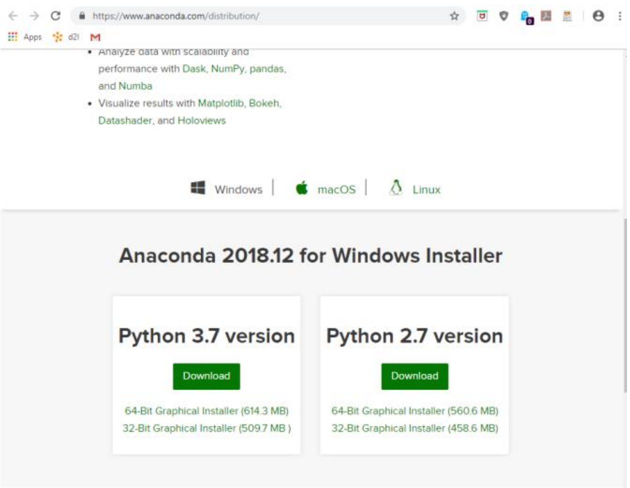
training set. It may increase accuracy to predict those keywords introduced later using training data from after their introduction date.

- Individual optimisation of algorithms. The algorithms were coarsely optimised and further investigation could improve upon this.
- Individual scaling. All keywords were scaled to 10%. Further investigation could investigate whether there is an optimum for each keyword.
- Stacking, random forest, other ensemble algorithms. Further investigation of ensemble algorithms could be considered.
- Including report title. Other text could be considered as input for predicting the keywords, such as the title, introduction or conclusion of the report.
- Other Deep Learning algorithms. Only the Gated Recurrent Unit (GRU) algorithm was implemented in this investigation. Further work could investigate other deep learning methods.

HOW-TO INSTALL NLP IN PYTHON

Installing Python with Spyder IDE

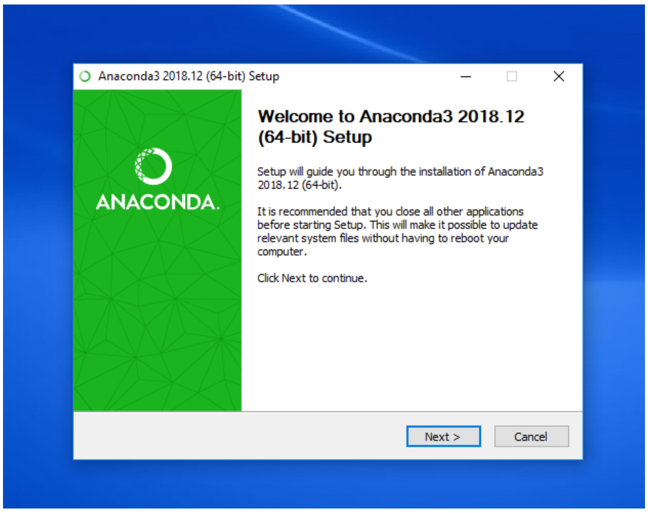
Spyder IDE (https://www.spyder-ide.org/) is a scientific integrated development environment for Python programming language. This allows the user to edit scripts, run them and interrogate variables with ease. The easiest way to download it is using the Anaconda distribution package which is a platform where many of the python packages have been bundled together for ease-of-install/access.



- Analyze data with scalability and performance with Dask, NumPy, pandas, and Numba
- Visualize results with Matplotlib, Bokeh, Datashader, and Holoviews

Anaconda 2018.12 for Windows Installer

Python 3.7 version	Python 2.7 version
Download	Download
64-Bit Graphical Installer (614.3 MB) 32-Bit Graphical Installer (509.7 MB)	64-Bit Graphical Installer (560.6 MB) 32-Bit Graphical Installer (458.6 MB)



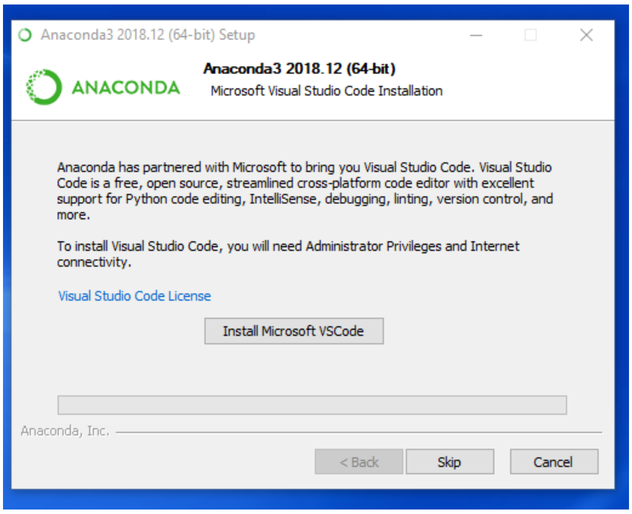
Welcome to Anaconda3 2018.12 (64-bit) Setup

Setup will guide you through the installation of Anaconda3 2018.12 (64-bit).

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Follow the installation guide, leaving all options as default. After installation on Windows machines, it will ask if you'd like to install Visual Studio Code. This is not required so 'skip'.



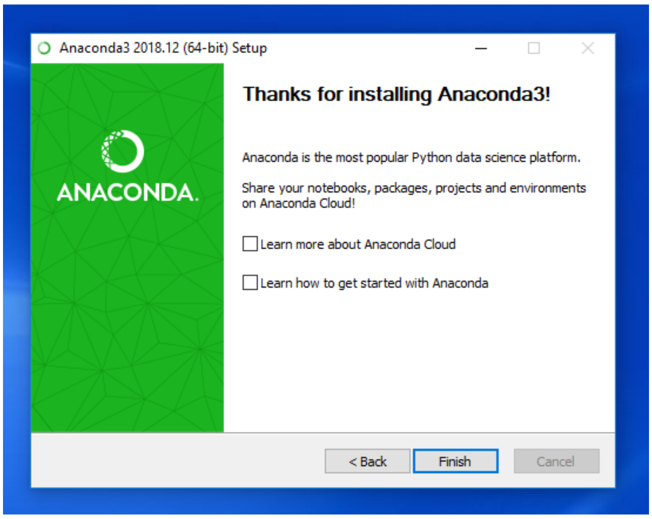
Anaconda3 2018.12 (64-bit)

Microsoft Visual Studio Code Installation

Anaconda has partnered with Microsoft to bring you Visual Studio Code. Visual Studio Code is a free, open source, streamlined cross-platform code editor with excellent support for Python code editing, IntelliSense, debugging, linting, version control, and more.

To install Visual Studio Code, you will need Administrator Privileges and Internet connectivity.

Visual Studio Code License



Thanks for installing Anaconda3!

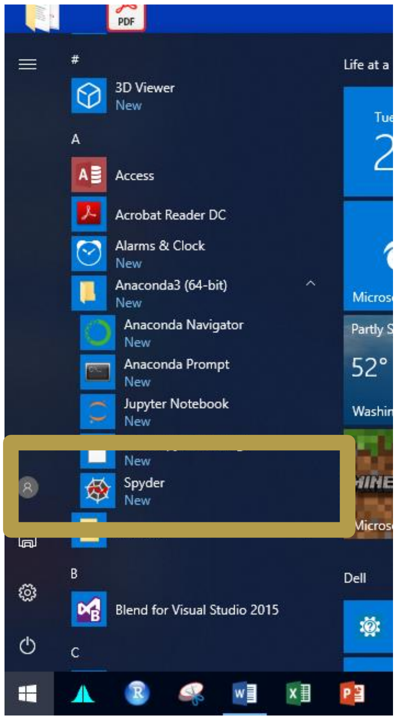
Anaconda is the most popular Python data science platform.

Share your notebooks, packages, projects and environments on Anaconda Cloud!

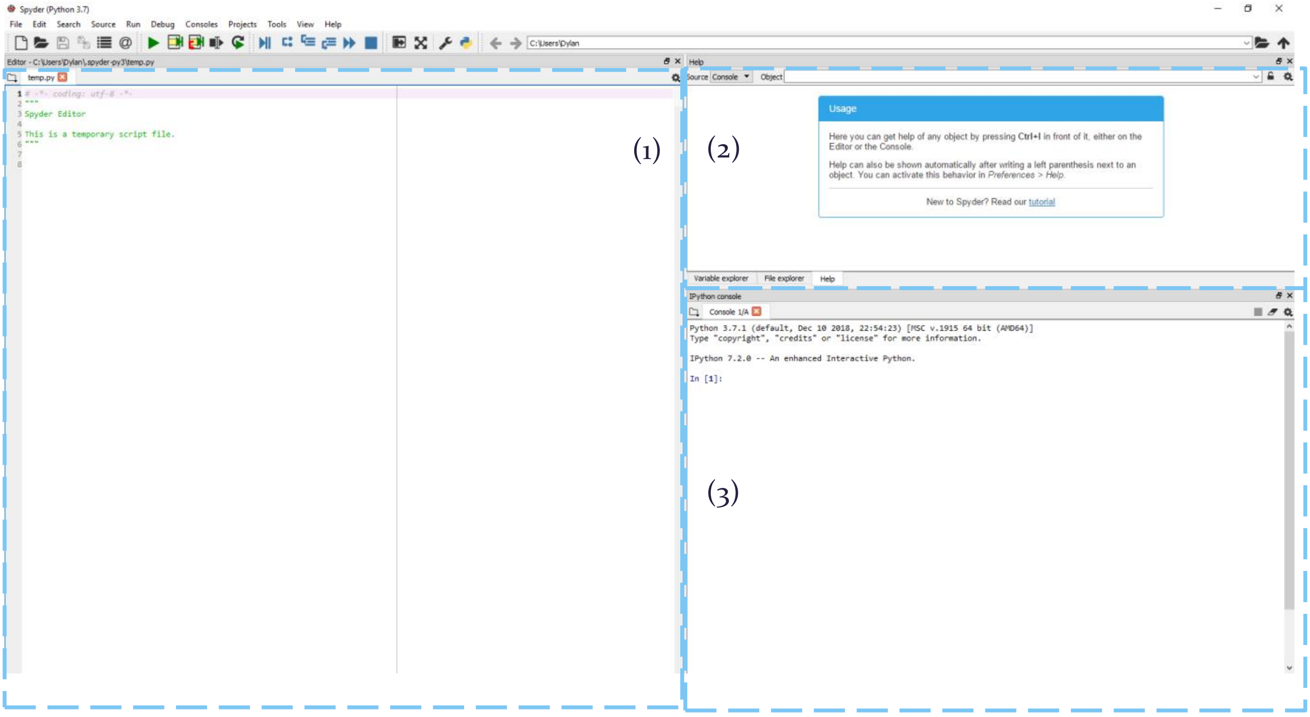
☐ Learn more about Anaconda Cloud

☐ Learn how to get started with Anaconda

After installation, the Anaconda file will appear in your startbar. Spyder will be in this folder and, to open Spyder, simply click on the icon.



Here are some quick notes on navigating Spyder. There are three main parts to the GUI: the script file editor (1), the IPython console (2) and the file/variable navigator (3). There are also some command icons along the toolbar at the top. The script editor allows you to view, edit and run Python scripts, while the console is where they are run and also allows you to run operations 'on-the-fly'. The file/variable editor allows you to view and interrogate your variables.



```
# -*- coding: utf-8 -*-
"""
Spyder Editor

This is a temporary script file.
"""
```

Here you can get help of any object by pressing **Ctrl+I** in front of it, either on the Editor or the Console.

Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in *Preferences > Help*.

New to Spyder? Read our tutorial

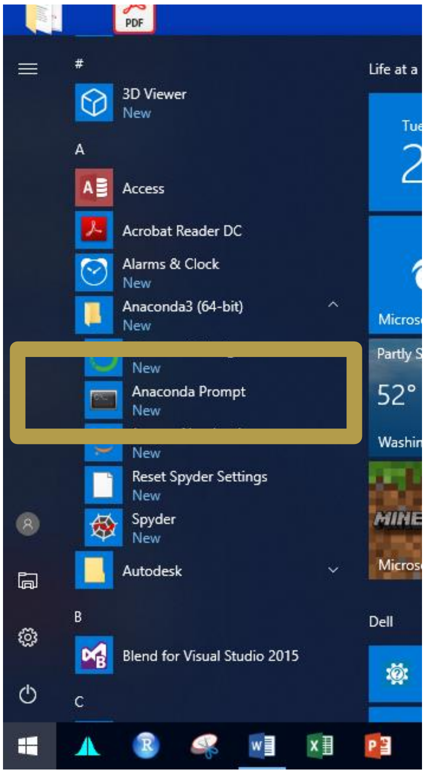
```
Python 3.7.1 (default, Dec 10 2018, 22:54:23) [MSC v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.2.0 -- An enhanced Interactive Python.

In [1]:
```

Installing extra packages

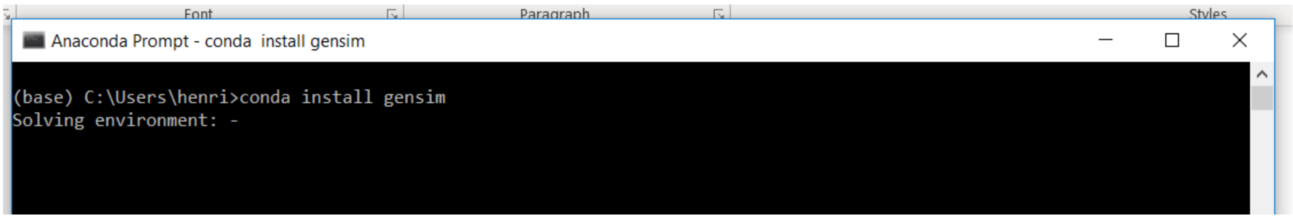
The Anaconda distribution package contains all the most commonly used Python packages; however, there are a few used in this project which are not included in this distribution. The most prominent is the gensim topic modeling ('gensim') package.

To install an additional package onto your computer, you must use the 'conda install' option. Do not try to install using 'pip' as suggested on many websites as it bypasses the Anaconda distribution and can break your installation.

To install packages, start by opening the Anaconda prompt which can be found on the start bar in the Anaconda3 file.

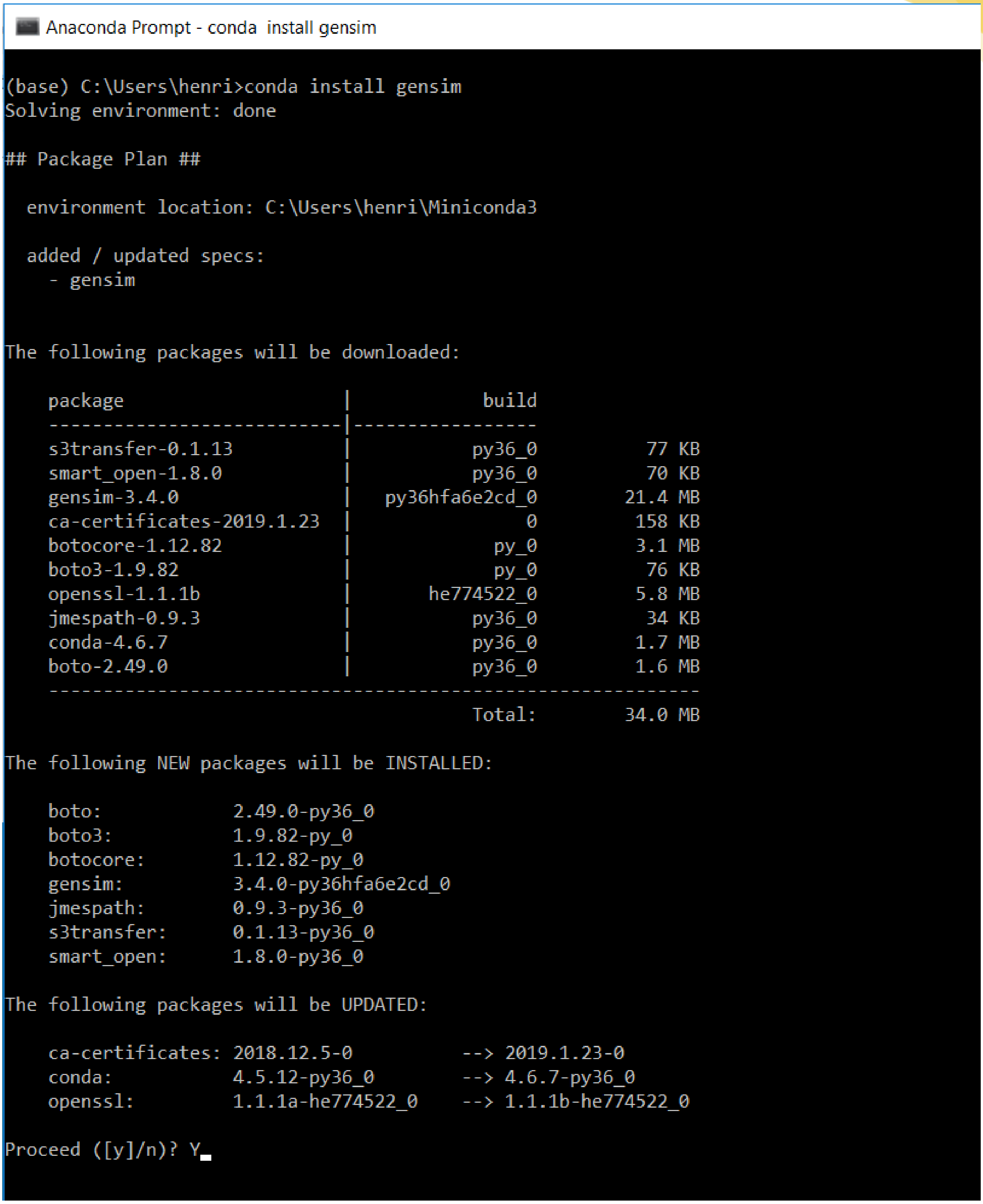


Next, write the cmd: 'conda install' + pkg name. E.g. 'conda install gensim'.



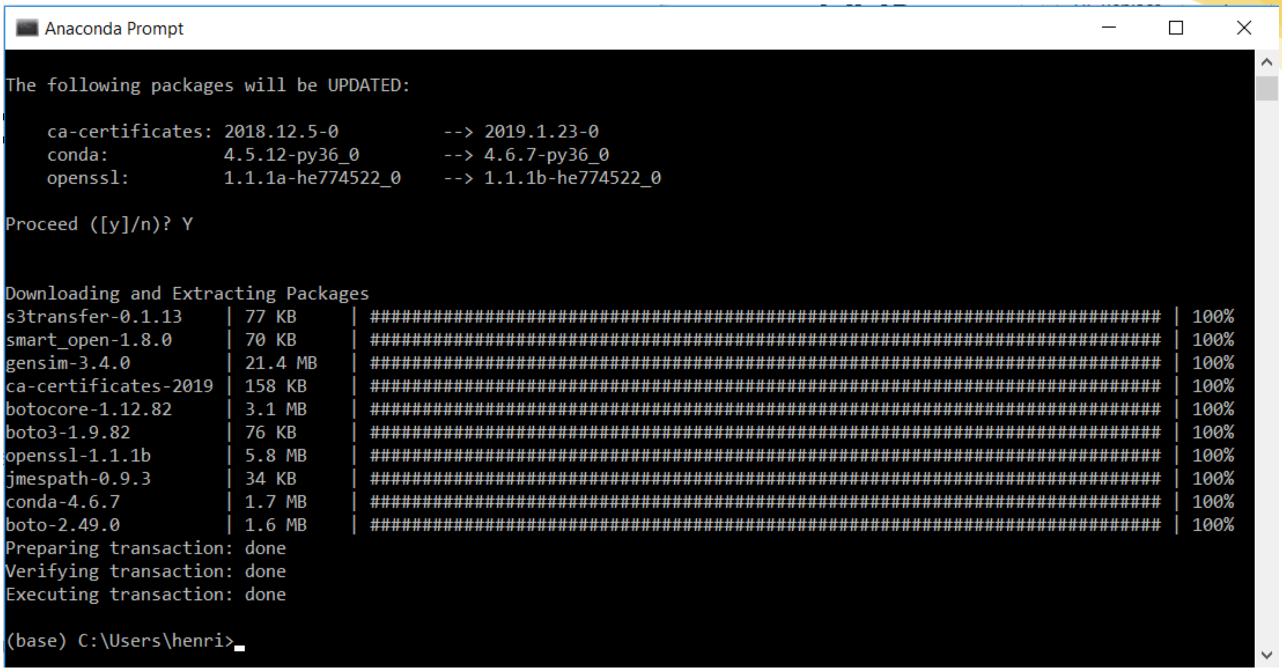
```
(base) C:\Users\henri>conda install gensim
Solving environment: -
```

The program will then do a system check. Select 'Y' to install the package:



```
(base) C:\Users\henri>conda install gensim
Solving environment: done

## Package Plan ##

  environment location: C:\Users\henri\Miniconda3

  added / updated specs:
    - gensim


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    s3transfer-0.1.13          |           py36_0          77 KB
    smart_open-1.8.0           |           py36_0          70 KB
    gensim-3.4.0               |   py36hfa6e2cd_0        21.4 MB
    ca-certificates-2019.1.23  |                0         158 KB
    botocore-1.12.82           |             py_0         3.1 MB
    boto3-1.9.82               |             py_0          76 KB
    openssl-1.1.1b             |       he774522_0         5.8 MB
    jmespath-0.9.3             |           py36_0          34 KB
    conda-4.6.7                |           py36_0         1.7 MB
    boto-2.49.0                |           py36_0         1.6 MB
    ------------------------------------------------------------
                                           Total:        34.0 MB

The following NEW packages will be INSTALLED:

    boto:            2.49.0-py36_0
    boto3:           1.9.82-py_0
    botocore:        1.12.82-py_0
    gensim:          3.4.0-py36hfa6e2cd_0
    jmespath:        0.9.3-py36_0
    s3transfer:      0.1.13-py36_0
    smart_open:      1.8.0-py36_0

The following packages will be UPDATED:

    ca-certificates: 2018.12.5-0          --> 2019.1.23-0
    conda:           4.5.12-py36_0        --> 4.6.7-py36_0
    openssl:         1.1.1a-he774522_0    --> 1.1.1b-he774522_0

Proceed ([y]/n)? Y
```

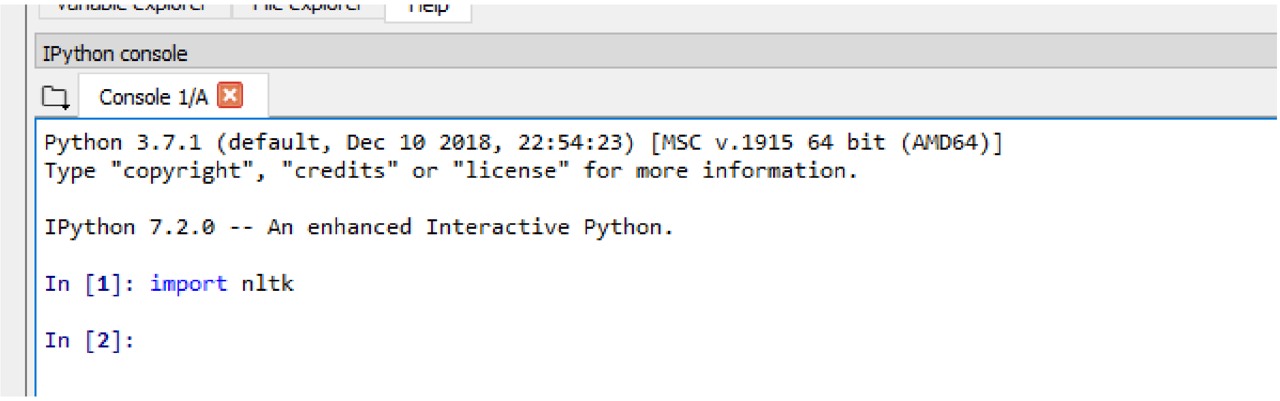
The package is now installed. You can check the installation by attempting to ‘import’ the package (see next section).



```
The following packages will be UPDATED:

    ca-certificates: 2018.12.5-0          --> 2019.1.23-0
    conda:           4.5.12-py36_0        --> 4.6.7-py36_0
    openssl:         1.1.1a-he774522_0    --> 1.1.1b-he774522_0

Proceed ([y]/n)? Y


Downloading and Extracting Packages
s3transfer-0.1.13    | 77 KB     | ############################################################################ | 100%
smart_open-1.8.0     | 70 KB     | ############################################################################ | 100%
gensim-3.4.0         | 21.4 MB   | ############################################################################ | 100%
ca-certificates-2019 | 158 KB    | ############################################################################ | 100%
botocore-1.12.82     | 3.1 MB    | ############################################################################ | 100%
boto3-1.9.82         | 76 KB     | ############################################################################ | 100%
openssl-1.1.1b       | 5.8 MB    | ############################################################################ | 100%
jmespath-0.9.3       | 34 KB     | ############################################################################ | 100%
conda-4.6.7          | 1.7 MB    | ############################################################################ | 100%
boto-2.49.0          | 1.6 MB    | ############################################################################ | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done

(base) C:\Users\henri>_
```

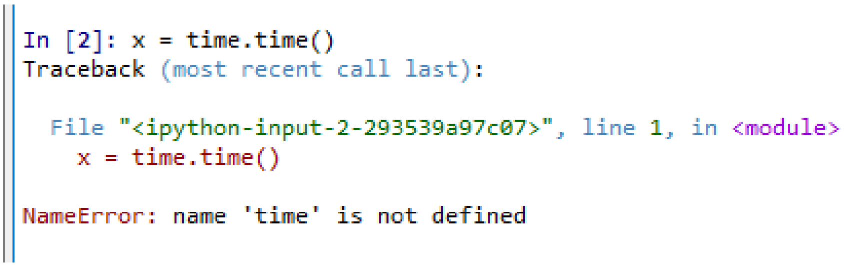
Importing packages

To run specific packages in Spyder/Python, it is necessary to 'import' them into the current session before use. This is simple. Use the command "import" and the package name, e.g. "import nltk", to import before use. An example is shown below:



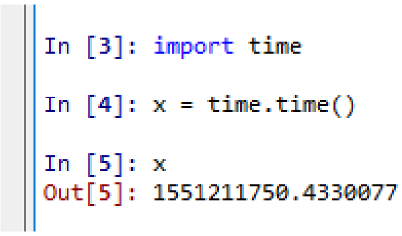
```
Python 3.7.1 (default, Dec 10 2018, 22:54:23) [MSC v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.2.0 -- An enhanced Interactive Python.

In [1]: import nltk

In [2]:
```

If you attempt to use a package and it hasn't been imported, the following error is given:



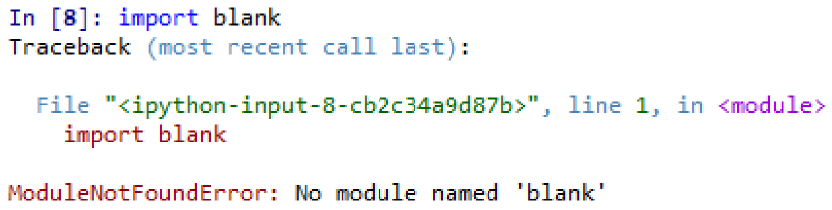
```
In [2]: x = time.time()
Traceback (most recent call last):

  File "<ipython-input-2-293539a97c07>", line 1, in <module>
    x = time.time()

NameError: name 'time' is not defined
```

vs



```
In [3]: import time

In [4]: x = time.time()

In [5]: x
Out[5]: 1551211750.4330077
```

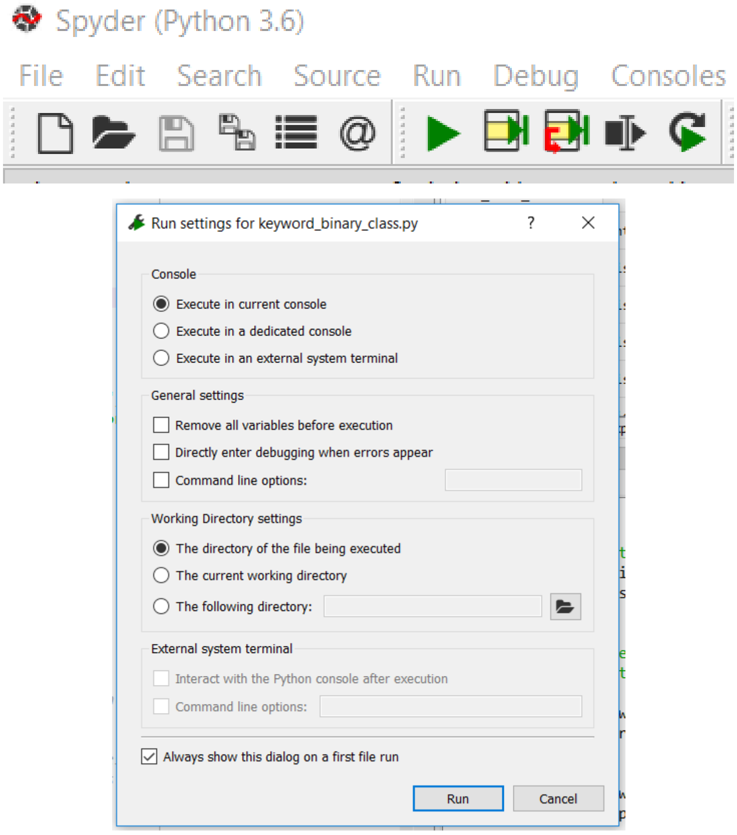
If you attempt to import a package and it hasn't been installed, the following error is given:



```
In [8]: import blank
Traceback (most recent call last):

  File "<ipython-input-8-cb2c34a9d87b>", line 1, in <module>
    import blank

ModuleNotFoundError: No module named 'blank'
```

Running programs

The easiest way to run programs from a .py file in Spyder IDE is to open the file using the 'open file' button on the top left of the console toolbar then press the green arrow button on the toolbar. If it is the first time you have run this program, it will then ask some generic questions about running it, leave these as the default (shown below) and select 'run'. The file will then run in the current session, as seen in the bottom right.



Run settings for keyword_binary_class.py

Console
- ● Execute in current console
- ○ Execute in a dedicated console
- ○ Execute in an external system terminal

General settings
- □ Remove all variables before execution
- □ Directly enter debugging when errors appear
- □ Command line options:

Working Directory settings
- ● The directory of the file being executed
- ○ The current working directory
- ○ The following directory:

External system terminal
- □ Interact with the Python console after execution
- □ Command line options:

☑ Always show this dialog on a first file run

Setting up GPU processing for deep learning

Follow instructions at https://medium.com/@ab9.bhatia/set-up-gpu-accelerated-tensorflow-keras-on-windows-10-with-anaconda-e71bfa9506d1

USER GUIDE FOR .PY SCRIPTS

This is a brief guide on running the keyword prediction in Spyder IDE for TWI staff using the files produced during this investigation.

Three scripts are included for TWI to run their own keyword prediction and continue this investigation, if desired. These are:

1. Keyword_classifier_v1.py
2. Embedding_pred_nopretrain_v1.py
3. compile_data_TWI.py

Additionally, the validationIDs.txt file is included to split out validation data.

The current Gated Recurrent Unit (GRU) deep learning implementation is slow and runs on CPU. To change to the fast running version (which uses GPU processing), edit in line 94 of 'Embedding_pred_nopretrain_v1.py'.

Set-up

1. Before opening Spyder IDE, place all four files along with the data file(s) the same folder. For this investigation, Weldasearch_records_20180521.txt was used. If new data are being used for prediction, they need to be in the same format (.txt extracted from Weldasearch). For prediction of keywords for new abstracts, these need to be in a separate .txt file in the same format but the keyword field will be blank.
2. Also in this folder, create a folder called **Data_processed** with two subfolders **outputs** and **temp**. It is important to create them exactly as named.
3. Start the Spyder IDE program.
4. Open Keyword_classifier_v1.py in Spyder.
5. The program is now ready to run and will open automatically when Spyder is next opened.

Compiling data

The first the program will compile input data for validation tasks. This requires user input if being compiled for the first time. Instructions are included throughout the process in the console; however, an overview of the steps taken is given below.

1. If the program has been run before, .csv file may already exist for the data in the Data_processed folder. The program checks for these and, if it finds them, it will ask:

   Q. **Both training_df.csv and validation_df.csv files exist. Do you wish to use existing files?**

   Answer by typing **Y** or **y** followed by 'enter' key to use the existing files. This will complete the compile phase for this run.

   If you wish to recompile the files or use new data, type **N** or **n** followed by 'enter' key. This will take the program to step 3.

2. If the files are not found, the program progress to step 3 after the following message:

   No training_df.csv exists, or it is in the wrong location. The data compiler will now run.

3. Next, the type of task is selected. There are two options: either a validation run, where a single file of Weldasearch data is used to obtain accuracy metrics, or keyword prediction for new abstracts. To select, the following question will be asked:

   Q. **Are you completing a validation run or predicting keywords for new data? Type: valid or new.**

   Answer by typing **valid** or **new** to select the type of task to compile the required data.

For a validation run, one data .txt input file (extracted from Weldasearch) is required, and this will be split into train, test and validation data to obtain accuracy results. The instructions for this are given in step 4 – 5.

To predict keyword for new abstracts, two data .txt input files are required. One which contains the new abstracts (in the .txt Weldasearch extraction format) and one which contains training data extracted from Weldasearch. This training file can be the same one used for the validation runs. The instructions for this task are given in step 4 – 5.

4. For validation runs, to obtain accuracy values, the data compile process will then run once after the following message is given:

   Compiling data for a validation run. Be sure to input the correct file name.

5. The compiler will then ask if you wish to use the default file which is currently the one used in this investigation: Weldasearch_records_20180521.txt.

   Q. **Is the file you wish to use named Weldasearch_records_20180521.txt? Y or N.**

   To use the named file, **Y** or **y** followed by 'enter' key. To edit the default file, you need to edit the file named in lines 196 and 204 of the compile_data_TWI.py file.

   To use a new file, type **N** or **n** followed by 'enter' key. The program will then instruct you to input the name of the file you wish to use by giving the message:

   Q. **Type filename including .txt extension. Ensure that the file is in the program file folder.**

   If the file is not found, or a spelling error occurs, then this instruction will loop with an additional error message:

   File not found. Check spelling and that the file is in the correct location. Remember to include the .txt extension.

6. For predicting keywords for new abstracts, the following messages will be printed:

   The data compile program will now run twice: once for the training data and once for the new abstracts.

   If a training_df.csv (compiled training data file) already exists, the following option is given:

   Q. **The training_df.csv file exist. Do you wish to use existing file? Y or N.**

   Answer by typing **Y** or **y** followed by 'enter' key to use the existing file. If you wish to recompile the file or use new data, type **N** or **n** followed by 'enter' key.

   Compiling training data. Be sure to input the correct file name.

   Step 5 will then commence for compiling the training data file. This will skip if the option to use the existing training_df.csv is selected.

   Compiling new abstract data. Be sure to input the correct file name.

   Step 5 will then commence for compiling the new abstract data file.

Running a validation task

These are steps to run a validation task to obtain accuracy results for the methods selected. This does not input new abstracts for keyword prediction.

1. Run Keyword_classifier_v1.py in Spyder using the green 'run' arrow.
2. A series of questions will then require user input in the console window:

   Q. **Would you like to use the 'bag-of-words' machine learning keyword predictions?**

   Answer by clicking in the console and typing **Y** or **y** followed by 'enter' key to run keywords prediction using the BoW + ML method. The current default methods are Gradient Boosting and SVM.

   The following input will then appear:

   Q. **Please input a number for the minimum keyword occurrence for ML runs**.

   Please type a number which will be the minimum number of abstracts assigned that keyword required in the training data to include said keyword in the prediction task. For example, if 100 is typed, there must be 100 abstracts with the keyword assigned to carry out a prediction task with that keyword. Any keywords assigned less than 100 times in the training dataset will be skipped.

   If this input is a low number, more keywords will be run so the runtime will be slower. As the accuracy for keyword prediction with sparse positive examples is low, it is suggested that this number be kept >1000.

   Q. **Would you like to use the deep learning keyword predictions?**

   Answer by typing **Y** or **y** followed by 'enter' key to run keywords prediction using word embedding method. If you do not wish to run the DL method, type **N** or **n** followed by 'enter' key. Note: only keywords which occur more in more than 0.6% of abstracts are predicted using DL methods.

3. Compile data for validation task. If the las task run was a validation task, the existing files can be used.
4. The program will then run prediction using the selected methods.

Running a new abstract keyword prediction task

These are steps to run a validation task to obtain accuracy results for the methods selected. This does not input new abstracts for keyword prediction.

1. Run Keyword_classifier_v1.py in Spyder using the green 'run' arrow.
2. A series of questions will then require user input in the console window:

   Q1. **Would you like to use the 'bag-of-words' machine learning keyword predictions?**

   Answer by clicking in the console and typing **Y** or **y** followed by 'enter' key to run keywords prediction using the BoW + ML method. The current default methods are Gradient Boosting and SVM. The following input will then appear:

   Q1.1 **Please input a number for the minimum keyword occurrence for ML runs**.

   Please type a number which will

   Q2. **Would you like to use the deep learning keyword predictions?**

   Answer by typing **Y** or **y** followed by 'enter' key to run keywords prediction using word embedding method. If you do not wish to run the DL method, type **N** or **n** followed by 'enter' key. Note: only keywords which occur more in more than 0.6% of abstracts are predicted using DL methods.
3. Compile data for new abstract task.
4. The program will then run prediction using the selected methods.

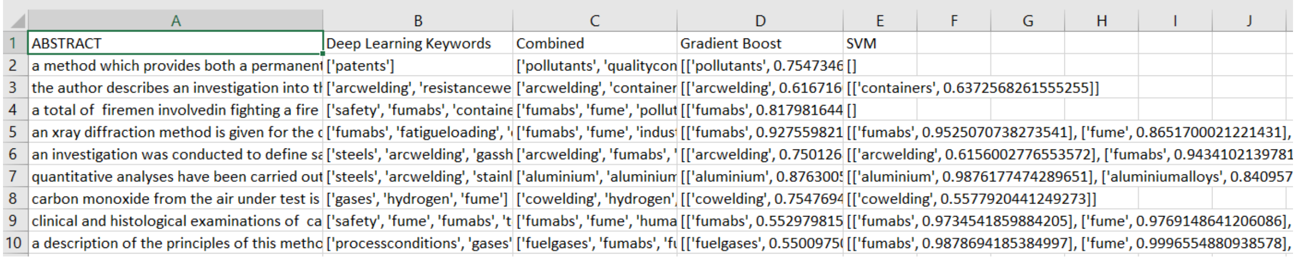
Output files

The following output files will be produced in the **output** folder, preceded by the date and time in their name:

a. 'Keywordresults.csv'
b. 'Validation-output.csv'
c. 'Output.csv'
d. Deep learning metrics
   i. 'Validation-nopretrain-results.csv'
   ii. 'Nopretrain-results.csv'

The 'keywordresults.csv' file contains the abstract text and keyword predictions for the new abstracts or validation data (depending on task). An example is seen below.

The first column contains the abstract text. The 'Deep Learning Keywords' column contains the keywords predicted by the GRU deep learning methods, while the 'Combined' column contains a joined list of those keywords predicted by all the BoW + ML methods combined. The subsequent columns contain the keywords predicted by each method with the column named after each method. Each cell in these columns contain a list of the keywords followed by their confidence score. High scores mean that the algorithm is confident that they have correctly assigned the keyword, scores closer to 0.5 mean that the algorithm is unsure.



	A	B	C	D	E	F	G	H	I	J
1	ABSTRACT	Deep Learning Keywords	Combined	Gradient Boost	SVM					
2	a method which provides both a permanent	['patents']	['pollutants', 'qualitycon	[['pollutants', 0.7547346	[]					
3	the author describes an investigation into tl	['arcwelding', 'resistancewe	['arcwelding', 'container	[['arcwelding', 0.616716	[['containers', 0.6372568261555255]]					
4	a total of firemen involvedin fighting a fire	['safety', 'fumabs', 'containe	['fumabs', 'fume', 'pollut	[['fumabs', 0.817981644	[]					
5	an xray diffraction method is given for the c	['fumabs', 'fatigueloading', '	['fumabs', 'fume', 'indus	[['fumabs', 0.927559821	[['fumabs', 0.9525070738273541], ['fume', 0.8651700021221431],					
6	an investigation was conducted to define sa	['steels', 'arcwelding', 'gassh	['arcwelding', 'fumabs', '	[['arcwelding', 0.750126	[['arcwelding', 0.6156002776553572], ['fumabs', 0.9434102139781					
7	quantitative analyses have been carried out	['steels', 'arcwelding', 'stainl	['aluminium', 'aluminium	[['aluminium', 0.876300	[['aluminium', 0.987617747428965l], ['aluminiumalloys', 0.840957					
8	carbon monoxide from the air under test is	['gases', 'hydrogen', 'fume']	['cowelding', 'hydrogen',	[['cowelding', 0.7547694	[['cowelding', 0.5577920441249273]]					
9	clinical and histological examinations of ca	['safety', 'fume', 'fumabs', 't	['fumabs', 'fume', 'huma	[['fumabs', 0.552979815	[['fumabs', 0.9734541859884205], ['fume', 0.9769148641206086],					
10	a description of the principles of this metho	['processconditions', 'gases'	['fuelgases', 'fumabs', 'fu	[['fuelgases', 0.550097	[['fumabs', 0.9878694185384997], ['fume', 0.9996554880938578],					

Files b-d contain accuracy metrics for the algorithms used. The program will also produce files 'validation-output.csv' and 'output.csv' after 500, 1000 and 1500 keywords predicted. As the program runs the most prevalent keywords first, this ensures that these results are saved first. These files contain the associated number to denote this e.g. 'validation-output-1000'.

For validation runs, 'validation-output.csv' and 'validation-nopretrain-results.csv' contain the reportable accuracies. For tuning any metrics/introducing new algorithms, 'output.csv' and 'nopretrain-results.csv' should be used. This ensures that the validation metrics are independent of any model tuning operation.

For predicting keywords for new abstracts, 'validation-output.csv' and 'validation-nopretrain-results should be disregarded.

Estimating Runtime

Using 16GB RAM CPU (Intel Core i7 7th Gen) and 4GB GPU (NVIDIA GTX 1050), the following timings were observed:

Method	**Runtime / keyword**
Naïve Bayes	1 sec
kNN	500 secs
Decision Tree	30 secs
Gradient Boost	180 secs
SVM	2 secs
Bagged SVM	15 secs
Gated Recurrent Unit, GRU (Fast – on GPU)	80 secs
GRU (on CPU)	220 secs
Set up time	15-20 mins / data compile
BoW vector set-up	5-10mins / run
BoW resampling Nb. this is will be long for small positive example keywords, and 0 secs for keywords >10%	Maximum 260 secs / keyword