TEXT MINING OF THE SECURITIES AND EXCHANGE COMMISSION FINANCIAL FILINGS OF PUBLICLY TRADED CONSTRUCTION FIRMS USING DEEP LEARNING TO IDENTIFY AND ASSESS RISK

A Dissertation
Presented to
The Academic Faculty

by

Yashovardhan Jallan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
August 2020

TEXT MINING OF THE SECURITIES AND EXCHANGE COMMISSION FINANCIAL FILINGS OF PUBLICLY TRADED CONSTRUCTION FIRMS USING DEEP LEARNING TO IDENTIFY AND ASSESS RISK

Approved by:

Dr. Baabak Ashuri
School of Building Construction & School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Eric Marks
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Iris Tien
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Caroline Clevenger
Department of Civil Engineering
University of Colorado Denver

Dr. Xinyi Song
School of Building Construction
Georgia Institute of Technology

Date Approved: June 10, 2020

To the Georgia Institute of Technology

ACKNOWLEDGEMENTS

I would first like to thank my doctoral adviser, Dr. Baabak Ashuri, who has had such a positive impact on my education and research odyssey. The lessons I have learned during our innumerous interactions will guide me for the rest of my career. Dr. Ashuri encouraged me to explore new ideas and inspired me through his valuable insight, advice and incredible supervision.

I also greatly appreciate the input and feedback I’ve received from my committee members; thank you to Dr. Iris Tien, Dr. Xinyi Song, Dr. Eric Marks and Dr. Caroline Clevenger. Your advice has greatly impacted this research and my holistic learning during my graduate studies, and for that I am deeply thankful.

I would also like to thank my all my course professors and lab group members, some of whom I have I had the pleasure to work with on various research projects during my stay at Georgia Tech. I made some amazing friends through shared courses and campus groups, which will stay with me going forward, and all of you have positively encouraged and molded my doctoral journey.

A special acknowledgment is dedicated to all the administrative staff in different offices across Georgia Tech. Your contribution can be easy to overlook, however the positive impact you have had on my life is immense, and I want to thank you for all you have done for me.

I am lucky to have a strong support from my wonderful family members and friends, who have always been there for me and helped me in achieving my goals and aspirations and have helped keep me motivated throughout this very rewarding journey.

Lastly, and most importantly, I would like to highlight the incalculable and unwavering love and guidance that I have received from my beloved parents. This dissertation is as much yours, as it is mine, and without you, none of this would be possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AI	Artificial Intelligence
BIM	Building Information Model
CIK	Central Index Key
CRMS	Construction Risk Management System
DBE	Disadvantaged Business Enterprise
DBIA	Design-Build Institute of America
DOT	Department of Transportation
DSC	Differing Site Conditions
EDGAR	Electronic Data Gathering, Analysis, and Retrieval
ENR	Engineering News Record
GDP	Gross Domestic Product
IoT	Internet of Things
IPD	Integrated Project delivery
IPO	Initial Public Offering
KNN	K-Nearest Neighbors
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MD&A	Management's Discussion and Analysis
MEP	Mechanical, Electrical and Plumbing
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NLU	Natural Language Understanding

P3	Public-Private Partnership
PART	Projective Adaptive Resonance Theory
POS	Parts of Speech
QA	Quality Assurance
RAM	Random-Access Memory
RFD	Risk Factor Disclosure
ROW	Right of Way
RT	Risk Type
SEC	Securities and Exchange Commission
SIC	Standard Industrial Classification
SME	Subject Matter Expert
SRTA	State Road and Tollway Authority
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
US	United States
VDC	Virtual Design and Construction

SUMMARY

Risk factors identification and mitigation has been a critical topic in the construction industry. Given the magnitude of large construction projects, the amount of capital involved, the inherent dangers of the work site, strict regulatory environment and ever-increasing business competition; it has become extremely important for the various construction enterprises to carefully understand and identify the different types of risks that affect their business and financial bottom line. A careful examination of industry-wide risk factors is useful for all the stakeholders involved. The current research creates a systematic methodology to identify and classify risk types affecting the construction industry. This research presents a new set of text mining methods to extract useful risk information from unstructured text data through carefully examining the financial filings of the publicly traded construction companies. Specifically, the 'Item 1A – Risk Factors' section of the 10-K reports filed by the public construction companies with the Securities and Exchange Commission (SEC), is leveraged as a previously unearthed source of insight into risk types affecting the industry. A structured procedure is developed to apply advancements from text mining, machine learning and natural language processing (NLP) in identifying the risk types from textual disclosures in the companies' filings. A state-of-the-art deep learning method based on word embedding algorithm developed by Facebook Artificial Intelligence (AI) Research, named FastText, is implemented, in order to identify risk patterns and classify the text into appropriate risk types. The methodology of the research is validated with the help of a structured survey of industry professionals along with evidence from literature.

Key findings show that operational and financial risks associated with doing business is most commonly disclosed in the risk disclosures filed by the publicly-traded construction firms. A steady monotonic increase is found in the average number of total risk disclosures per file from 2006 to 2018. Over the same period, a growth is seen in the proportion of technology risks, reputation/intangible assets risks, financial markets risk and third-party risks. The primary contributions of this research are: (a) development of a new methodology which serves as a risk thermometer for identification and quantification of risk at an individual company level, sub-industry level, and the overall industry level; and (b) minimization of any existing information asymmetry in risk studies by utilization of a source of data that have not been previously used by construction researchers. It is anticipated that the developed methodology and its results can be used by: (i) publicly-traded construction companies to understand risks affecting themselves and their peers; (ii) surety bond companies and insurance providers to supplement their risk pricing models; and (iii) equity investors and capital financial institutions to make more informed risk-based decisions for their investments in the construction business.

CHAPTER 1. INTRODUCTION

1.1 Risk Analysis in Construction Industry and Existing Opportunities

Risk analysis in the construction industry has been a topic of prime interest for both research and professional communities. Risk is directly tied to the financial success of a construction firm but in addition to that, it is also tied to the success in terms of quality of the project delivered, the safety of the workers involved, protection from claims and liabilities, schedule delay and extension avoidance, and compliance with the several laws that govern the industry. As a large industry which is a major component of the Gross Domestic Product (GDP) of United States, it is extremely important to keep up with the dynamic changes in the impact and influence of various risk factors that are widespread in the construction industry.

Success of any large construction project is often dependent on various stakeholders. A stakeholder is any entity that has an interest in the process or outcome of a construction project. In a typical project, there are various different entities that have a stake in the decisions and implications of the projects, such as, the owner (client), the main contractor, materials supplier, equipment manufacturer, designers, subcontractors, employees in any capacity in the project, local authorities, the end users, professional certifying bodies, local residents, local business owners, politicians, lobby groups, investors, financiers, insurance companies, legal enterprises etc. This is a representative list as the actual stakeholders for a specific construction project largely depend on the specifics of the project. In such a setting, all stakeholders have an interest to carefully examine the risk profiles and reward expectations of all kinds of construction companies working on the project. The construction companies themselves have an incentive to keep abreast with risk factors of other companies and take notice of their competition and the type of risks that they are facing. The current work addresses some of these existing opportunities.

As per Figure 1, the author demarcates different levels of analyzing risk in the construction industry as Task, Project, Program, Enterprise, Sub-Industry and Industry levels. These levels are set based on the scope and granularity of risk assessment. With risk identification being such a crucial factor in the construction industry, researchers have used various techniques and adopted a myriad of methodologies to look at different types of issues within the industry. Each level of risk assessment presents unique challenges. Careful examination and understanding of all the potential adverse eventualities are important.

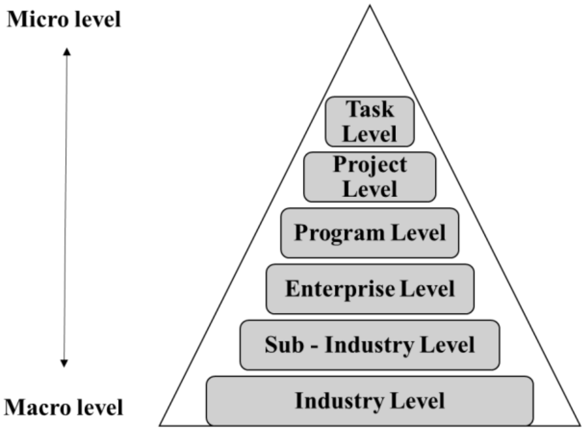


Figure 1 - Different levels of analyzing risk in the construction industry

Table 1 presents brief descriptions of the different levels of analyzing risk in the construction industry, with some relevant examples to further elucidate the basic premise.

Table 1 - Brief description of levels of risk analysis in construction industry with examples

Risk Level	Brief Description	Examples
Task Level	Risks associated with individual tasks in a construction project.	Ex: Quality Assurance, Scheduling, Right - of - way (ROW) acquisitions etc.
Project Level	Risks that affect the overall project.	Ex: Selection of prime contractor, competence/experience of the head project manager etc.
Program Level	Risks associated with a group of similar projects.	Ex: All Design-Build projects for a State DOT, all bridge projects for a prime contractor etc.
Enterprise Level	Risks that are associated with running an organization.	Ex: Leadership by the chief executives, vision and direction for the entire company, financial condition of the organization etc.
Sub - Industry level	Risks associated with all the companies that provide a similar type of service.	Ex: all electrical work sub-contractors, MEP sub-contractors, design firms etc.
Industry level	Risks associated with the overall construction industry.	Ex: Federal spending in infrastructure, material prices, competition from other countries, slowdown in economy etc.

1.1.1 Literature review of Task Level Risk studies

When the risk level is granular and scope is smaller, the researchers can employ more elaborate and detailed analysis. Some examples of Task level studies are, Gad and Shane (2017) developed a model focusing on culture-risk-trust in selection of dispute resolution methods for international construction contracts. They employed a Delphi technique surveying expert views on the factors and recommended different dispute resolution strategies depending on project specifications. Wang et al. (2017) assessed the work-related risk factors on lower back disorders among the roofing workers. Ashuri et al. (2018) compared the risk-based methods for quality assurance (QA) in highway projects of State DOTs, and presented a survey of latest approaches to identify and classify risk in the QA program of State DOTs. Alomari et al. (2018) surveyed construction safety professors and practicing safety engineers to investigate the extent and impact of different factors on worker safety risk.

1.1.2 Literature review of Project Level Risk studies

Various studies have been conducted by researchers to look at Project level risk. Al-Bahar and Crandall (1990) first introduced a risk model entitled construction risk management system (CRMS) to help contractors identify project risks and systematically analyze and manage them. They used Monte Carlo simulation techniques to analyze and evaluate project risks and suggested strategies to avoid, transfer and reduce risk. Creedy et al. (2010) evaluated risk factors that lead to cost overruns in delivering highway construction projects. They examined owner risk variables using multivariate regression analysis and found reciprocal relationship between project budget size and percentage of cost overruns. Tran and Molenaar (2015) developed a risk-based model analyzing project cost, risk and uncertainty in determining the most suitable delivery method in highway projects. Jarkas and Haupt (2015) published a study that identified and explored the prevalent allocation response trends of the major construction risk factors considered by general contractors in Qatar. They employed a contractor survey-based approach to relatively rank a list of 37 potential risk factors.

1.1.3 Literature review of Program Level Risk studies

Coming to Program level of risk analysis, Touran (2014) provided a mathematical framework to model cost uncertainty and escalation for a portfolio of large infrastructure projects with multi-year duration. The study considered randomness of cost and escalation factor per project. Zhao et al. (2016) prepared a list of 28 risk factors grouping them into 11 classes, that plagued the green building projects in Singapore. The study implemented a survey methodology from project managers of construction projects and found that risks associated with cost estimation and cost overrun were paramount. Ashuri et. al. (2018) studied risk identification strategies to enhance the delivery of highway projects. They conducted interviews with subject-matter experts in different functional offices of State DOTs and identified important risk types.

1.1.4 Gaps and Opportunities in Macro Levels of Risk studies

On moving to more macro levels of analyzing risk, researchers note that the larger scope of the challenge poses a problem with respect to designing a methodology that accurately captures a broad-based view of different risk types and evaluating the implications and severity of the identified risks across an organization. Hallowell et al. (2013) conducted a detailed analysis to come up with risk identification strategies for State DOTs at an enterprise level and described the benefits of this approach as: -

- Enterprise risk management allows common risks that have traditionally been managed at the project level to be more efficiently and consistently managed across the enterprise.

- It facilitates the inclusion of risk management into financial analyses and other primary organizational functions at various levels of the organization.
- This integration yields a positive return on investment because risks can be managed across the organization, and the downsides of single project risk management, such as overestimating individual project contingencies, can be avoided.
- The benefits of this approach include, but are not limited to: more efficient use of vital resources; the ability to evaluate risk interdependence and manage strategic risks; improved financial stability; and the development of a culture of risk management.
- Specifically, this approach avoids risks from being managed multiple times by different functions within the department.

Even with the numerous benefits listed above, analyzing risks on an enterprise level, sub-industry level or industry level can often be a herculean task, just by the size and scope of the problem at hand. Typically, researchers run into major challenges, such as lack of a consistent data source that captures the information that would provide a factually correct summary of the risks faced by the construction companies. Hypothetically, even if a researcher is able to sample data from a significantly large pool of construction companies, the usual methods of manually analyzing the survey results to make an apple to apple comparison can be a daunting task. The time and resources that such a study would require is a major limiting factor. Additionally, performing a uniform comparison across the companies is challenging as subjectivity can be introduced in form of recency biases, inadvertent errors can occur due to neglect and limitations of human capacity to comprehend hundreds of separate surveys and/or a multitude of granular statistical measures together. On top of that, to keep up with the dynamic nature of risks in the construction industry, conducting a survey-based methodology. Furthermore, such a research may still be theoretically possible for government agencies, such as State DOTs, which are funded by taxpayers' dollars, and thus are responsive to the survey questions of researchers. However, publicly traded construction companies may not have a clear incentive or obligation to divulge their business proceedings. In fact, they can sometimes be incentivized to keep their trade secrets within the bounds of the company, making it much harder for researchers to gather detailed information and data from such companies. This research aims to overcome these great challenges, in order to provide significant benefits of risk identification at the macro level for the construction industry. For the first time, information embedded in the 'Item 1A - Risk Factors' section of the 10-K SEC filings of publicly traded construction companies is used as an untapped source of data to identify and assess the risk factors at the macro level. 10-K SEC filings as professionally audited sources provide unique opportunities to conduct a structured, data-driven and systematic analysis of risk for construction firms at macro level.

1.2 The SEC 10-K filings

The U.S. Securities and Exchange Commission (SEC) is an independent agency of the United States federal government whose primary responsibilities include enforcing the federal securities laws, proposing securities rules, and regulating the securities industry including the nation's stock exchanges and all the public companies (SEC.gov | What We Do n.d.). It also plays the role of ensuring that all publicly traded companies are completely transparent in their business and financial dealings, and requires these companies to submit various types of information through a system of financial filings and reporting at regular intervals. These financial filings are made available by the SEC to the public through the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, which is a database that is available for free and open access on the internet (SEC.gov | About EDGAR n.d.). Of the various types of financial filings mandated by the SEC, a comprehensive annual report known as '10-K Report' is typically the most detailed and scrutinized document that is filed by each public company within 90 days after the end of its fiscal year (SEC.gov | Form 10-K n.d.). The SEC requires that the 10-K report be professionally audited. It is intended as the most important resource for potential investors to understand the financial affairs of a company and plan their investment decisions accordingly.

1.2.1 Overview of Item 1A - Risk Factors section of the 10-K filings

Out of all the different sections in a typical 10-K file, the focus of the present research is on the 'Item 1A - Risk Factors' section. In 2005, the SEC began requiring an 'Item 1A - Risk Factors' to be a separate section of disclosure in the 10-K reports (SEC.gov | How to Read a 10-K n.d.). In this section the company discloses the most significant risks that could harm its business. It can range from risks inherent to a specific industry, like an offshore oil driller's risk of losses from a major accident and oil spill, to something that is broader, like a popular brand losing favor with consumers. Companies are required to use plain English in describing these risk factors, avoiding overly technical jargon that is difficult for a layperson to follow.

Table 2 shows an example of three risk factors listed by Fluor Corporation in its 10-K form of 2018. These three risk factor disclosures are good representatives of how different companies use their own way to describe some of the risks that they consider to be important and would like potential investors in their company to be aware of. There is no standard way to describe a particular risk disclosure (i.e., there is no standard classification of risk factors that companies have to follow in reporting their own risk disclosures).

Table 2 - Three sample risk factors in 'Item 1A - Risk Factors' section of Fluor Corp's 10-K Form of 2018

Three sample risk factors
We may experience reduced profits or losses under contracts if costs increase above estimates.
Intense competition in the global engineering, procurement and construction industry could reduce our market share and profits.
We are dependent upon suppliers and subcontractors to complete many of our contracts.

1.2.2 Use of the construction industry 10-K filings in the current research

The heart of this research effort is to narrow down and study these financial filings of the universe of publicly traded companies which fall within the construction and infrastructure industry classification. The focal point of the study is to discover, identify and quantify the risk factors that have governed the construction industry since the beginning of risk factors data availability starting from the year 2006. The embedded information in 10-K forms has never been utilized in the construction research to identify the risk factors impacting different firms and the overall construction industry sector. This research provides the first application to utilize advances in natural language processing

(NLP) and deep learning to identify, categorize and quantify the risk factors affecting the financial well-beings on the construction industry.

1.3 Natural Language Processing (NLP) and Deep Learning

Text data, by nature, is comparatively harder to work with on computers because machines understand the language of numbers. Building intelligent systems which can interpret and understand the free-flowing natural language like humans is a non-trivial task. Natural Language Processing (NLP) is the sub-field in computer science and artificial intelligence (AI) that facilitates the programming of computers to process and understand natural language data. Deep learning is a sub-field of Machine Learning that specifically uses artificial neural networks to understand complex patterns in unstructured text data to enable data analytics and decision making.

1.3.1 Literature review of Text Mining applications in the Construction Industry

Text mining, as a technique has witnessed increased usage in the construction industry research with several successful implementations as summarized in Table 3. Some of the earliest work in this niche was done by Caldas and Soibelman (2003), when they created an automated hierarchical document classification for managing construction documents. Tixier et. al. (2016) developed a content analysis tool for implementation in construction safety, to identify the causes and outcomes from injury reports. Le and Jeong (2017) applied an NLP-based methodology to extract and examine semantic relations of important data terms used in design manuals of different highway agencies. Most recently, Zhang and Ashuri (2018) mined building information model (BIM) log files to measure design productivity. Mahfouz et. al. (2018) conducted a study to identify the latent legal knowledge in differing site conditions (DSC) litigation cases. Jallan et. al (2019) developed a word-frequency model to study the important construction defect cases and applied topic modeling to identify important themes.

In other applications, Williams and Gong (2014) developed a procedure to combine textual description of a construction project with numerical data to predict the level of cost overrun using datamining algorithms. Williams and Betak (2016) used railroad accident data from federal railroad administration databases, to identify themes in railroad equipment accidents using text mining and text visualization. Moon et. al. (2018) developed a system named UNI tacit to automatically collect data and retrieve information from construction related news articles, reports and cases. Marzouk and Enaba (2019) developed a dynamic text analytics for contract and correspondence (DTA-CC) model to tackle current research gap, by developing a descriptive text analytical model to monitor correspondence sentiment and communication nature. All these studies have had success in implementing text mining techniques in the realm of construction research and has positively reinforced the hypotheses and methodology of the present research.

Table 3 - Review of text mining applications in construction research

Author(s)	Research Contribution	Techniques applied
Caldas and Soibelman (2003)	Created an automated hierarchical document classification for managing construction documents.	Classification algorithms like Naïve Bayes, KNN, SVM etc.
Williams and Gong (2014)	Developed a procedure combining textual description of a construction project with numerical data to predict level of cost overrun using datamining algorithms.	Transformation of text to numeric vectors, SVD

Table 3 Continued

Tixier et. al (2016)	Developed a content analysis tool for implementation in construction safety, to identify the causes and outcomes from injury reports.	Topic Modeling, Rule-based NLP
Williams and Betak (2016)	Used railroad accident data from federal railroad administration databases, to identify themes in railroad equipment accidents using text mining and visualization.	Topic Modeling, Data Visualization
Yarmohammadi et. al (2016)	Extracted implicit process information from design log data by implementing a sequential pattern mining approach.	Text parsing, Text cleaning and organizing.
Le and Jeong (2017)	Extract and examine semantic relations of important data terms used in design manuals of different highway agencies	Information Extraction (IE), Information Retrieval (IR), Rule-based NLP
Zhang et. al (2018)	Mined building information model (BIM) log files to measure design productivity.	Pattern Identification, Social Network Modeling, Relation Determination.
Mahfouz et. al. (2018)	Conducted a study to identify the latent legal knowledge in differing site condition (DSC) litigation cases.	Naïve Bayes, Decision Tree, and PART
Moon et. al. (2018)	Developed a system named UNI tacit to automatically collect data and retrieve information from construction related news articles, reports and cases.	Web Crawling, POS tagging, Word Cloud Visualization
Jallan et. al (2019)	Developed a word-frequency model using text mining to automatically identify and analyze construction defect cases, applied Topic Modeling to study important themes.	Text parsing, Word Frequency Analysis, Topic Modeling, LDA
Marzouk and Enaba (2019)	Developed a dynamic text analytics for contract and correspondence model, a descriptive text analytical model to monitor sentiment and communication.	Word Frequency Analysis, Sentiment Analysis, Clustering

1.3.2 Literature review of research conducted on 10-K textual filings

Text information embedded in 10-K filings of construction firms has never been utilized in the construction research. In the broader area of general finance and accounting, 10-K filings have been used to research risk. For example, Campbell et al. (2014) examined the information content of the newly created Item - 1A section for all the 10-K files between 2005-2009. They used a manual procedure and a predefined dictionary to quantify five risk types in 10-K forms: idiosyncratic, systematic, financial, tax, and litigation risks. Mirakur (2011) randomly sampled 122 firms and downloaded their 10-K files submitted in the year 2009, and manually categorized the risk factors into 29 risk types. Huang and Li (2011) used their subject matter expertise in financial accounting and read hundreds of annual reports, in order to come up with the 25 risk categories. Then, they implemented a supervised learning algorithm to place risk factors reported in Item 1A of the 10-K forms into those predefined risk types. Miihkinen (2013) conducted a study in Finland which examined the risk disclosures of Finnish firms during 2006-2009, and demonstrated that information asymmetry decreases with the quality of risk disclosure. The process involved a detailed manual content analysis of the risk disclosure documents. Chin and Moffit (2018) examined the risk factors in 10-K reports to understand the significance of order in which the risk factors were presented in the report. To study this hypothesis, they focused on the firm’s disclosures on credit risk and associated them with the firm’s credit rating and its bond spreads. The study employed keyword-based identification techniques to find the relevant disclosures on credit risk.

The studies presented above require a significant amount of human effort in terms of manual labor of reading though large numbers of files and identifying risk factors. Bao and

Datta (2014) were first researchers to explore the applicability of unsupervised machine learning techniques to classify the risk factors for a large number of 10-K files. They collected the entire database of 10-K filings for the period 2006-2009 and applied a variation of topic modeling algorithm called sent-LDA and generated 30 risk groups. Topic modeling algorithms fall within the domain of traditional count-based text mining methods which are known to lose information like the semantics, structure, sequence and context of the words in a textual format. While being an excellent first foray of unsupervised learning applied to risk factor disclosures, the researchers describe challenges in interpretation of the topics generated, as it can be very open-ended and thus causing issues in clearly delineating different risk groups. In the last decade, exponential advancements have been made in text mining algorithms and research, which are found to outperform the traditional count-based methods. The researchers also recommend that more robust methods should be applied to the problem for future research.

The present research creates its own niche by using the rich SEC 10-K filings data source to provide a custom-made solution specifically for the domain of construction industry, which has not been done before. State-of-the-art natural language processing and deep learning approach is implemented in order to overcome challenges associated with traditional text mining methods, to develop a robust and cutting-edge model for identifying, classifying, and quantifying the risk factors in the construction industry.

1.4 Research Contributions

The contributions of this research are: (a) development of a new methodology for risk identification and quantification at an individual construction company level, groups of companies within the same sub-industry level and the overall industry level; and (b) minimization of any existing information asymmetry by utilization of an unearthed source of data that have not been previously used by construction researchers (i.e., the SEC filings) to systematically discover the risks affecting the financial bottom line of construction firms.

It is anticipated that the public construction companies which file 10-K reports can gain good insight from the study by understanding the behavior of their peers and the industry as a whole. Private construction companies who are not mandated to file risk disclosures, can also benefit from the findings of this research and compare their internal risk assessment with the identified risks of the publicly-traded companies. Quality of the 10-K filing for these companies can improve because of the scrutiny placed by the existing research and any future research that the present research may inspire. The investors looking to invest in construction companies, can be benefitted from the research as well if they use the trends and patterns in risk factors and relate it to a positive return on investment. The pricing of risk into construction contracts can be made more efficient and cost-effective as the research can add previously unknown insights to the process. Insurance companies are on the lookout to be able to attribute appropriate importance to each risk factor and this study can help them price their insurance products to make it more favorable and competitive for potential construction owners and contractors. The hedging instruments used by the construction companies to safeguard against the fluctuations in various risks (like oil prices etc.) can be modelled more efficiently. Overall, it is expected that the research will add meaningful contributions to all the stakeholders involved in the construction industry and research, by providing an additional way of examination of industry wide risk factors.

Another important benefit of the present study is that it uses a database which is a continuous and growing source of information and the methodology is automated and quick, thus it enables the results to be quickly reproduced in the future and is expected to only get more detailed and richer with time. The research also allows an apple to apple comparison among companies (as some other survey methods can introduce biases due to interpretation and subjectivity but the SEC official guidelines are very uniform and strict so all public companies need to adhere to the guidelines.)

1.5 Organization of the Dissertation

This dissertation is organized into six chapters. It also includes two appendices containing supporting information for this research. Chapter 1 introduces the important themes of the existing landscape and the inter disciplinary nature of the research. Chapter 2 enumerates the main objectives, scope and hypotheses developed for this research. Chapter 3 introduces and elaborates on the SEC 10-K filings data source and how the present research uses the data in order to analyze and examine it. Chapter 4 demonstrates how the risk types are laid out, which is followed by an elaborate description of the text mining and machine learning techniques used for text classification. Chapter 5 presents a discussion of the results obtained after performing the research and goes into the validation procedure. Lastly, the conclusions, limitations, and future work recommendations of this research are discussed in Chapter 6.

CHAPTER 2. OBJECTIVES, SCOPE AND HYPOTHESES

In order to understand the goal, purpose and methodology of this research, it is important to define the overall objectives, scope and hypotheses. In this chapter, each of these research components are discussed.

2.1 Objectives

The major objective of the present research was to develop a methodology to discover, identify and quantify the risk factors prevalent in the construction industry, using the SEC database of 10-K reports for publicly traded construction companies. This overall problem statement is approached from a data science, natural language processing and machine learning point of view. To achieve this broad objective, the secondary objectives are listed here as follows: -

- Build a method that uses the SEC EDGAR database to download and assemble the 10-K filings for the publicly traded construction companies.
- To develop a process of programmatically extracting the ‘Item 1A - Risk Factors’ section from the collection of 10-K files.
- To develop and implement an in-depth methodology applying the advances and knowledge of text mining, natural language processing and machine learning to successfully discover and identify the various risk factors.
- To develop a methodology to summarize and quantify the identified risk types at a firm, sub-industry and industry level.
- To compare the research results with existing literature and knowledge.

- To validate the findings of text mining using human subject validation with the help of including participation of subject matter experts (SMEs).

2.2 Scope

The research will have focused on the 10-K filings of the publicly traded construction companies which are made available by the SEC EDGAR database. Within the 10-K filings of the construction companies, the focus is on the ‘Item 1A - Risk Factors’ section to be able to discover, identify and quantify the risk factors. The geographical scope of this research includes all the public construction companies which are listed as a company registered within the geographical boundary of the United States. The time period for data availability and analysis is 2006-2019.

As far the various technical methods employed are concerned, the data collection and manipulation are done using web scraping, natural language processing and text mining techniques. While the risk identification and classification from underlying textual data is achieved using data science, deep learning and text similarity measures. All these methods and analysis were applied considering the knowledge and understanding of the domain of the construction industry and its practices.

2.3 Research Hypothesis

Based on the review of literature concerning risk studies in construction industry, the success of financial research in the study of SEC filings of companies, and the implementation and success of text mining techniques in broader construction research; coupled with the first-hand experience of the researcher under the advisement of the doctoral adviser in study of these broad areas and implementation of similar techniques in past research, course projects and industry experience; the following hypotheses were developed :-

- The ‘Item 1A - Risk Factors’ in 10-K filings for publicly traded construction companies in the United States, can be successfully downloaded, extracted and processed for textual analysis.
- An automated methodology can be developed to synthesize the risk factors data in textual format that can be classified into easily interpretable and usable risk types.
- An approach can be created to summarize the identified risk types at a firm, sub-industry and industry levels.
- The processes to identify and quantify the risk factors can be demonstrably validated by comparing with existing knowledge and literature, and concurrence with subject matter experts.
- The research will lead to a development of a new methodology for risk identification and quantification that minimizes any existing information asymmetry and leads to practical implications for academia and industry.

2.4 Overall Flowchart of the Research Process

Figure 2 presents the overall flowchart that captures the methodology which was conducted for the present research. The methodology can be segmented into the following steps: -

a) Data Collection - which involves identifying the correct 10-K files and downloading them from the SEC EDGAR database

b) Data Manipulation - that covers extracting the risk factor disclosures and applying pre-processing and text cleaning operations to make them ready for text mining analysis

c) Defining the Risk Types - identifying the important risk types that need to be extracted

d) Deep Learning - includes the application of the FastText algorithm to convert the text into word vectors

e) Text Classification - which makes use of the cosine similarity measure to map the risk factor disclosures to the target risk types, and

f) Visualization - that involves creating Tableau dashboard to effectively visualize the results and risk patterns for discussion.

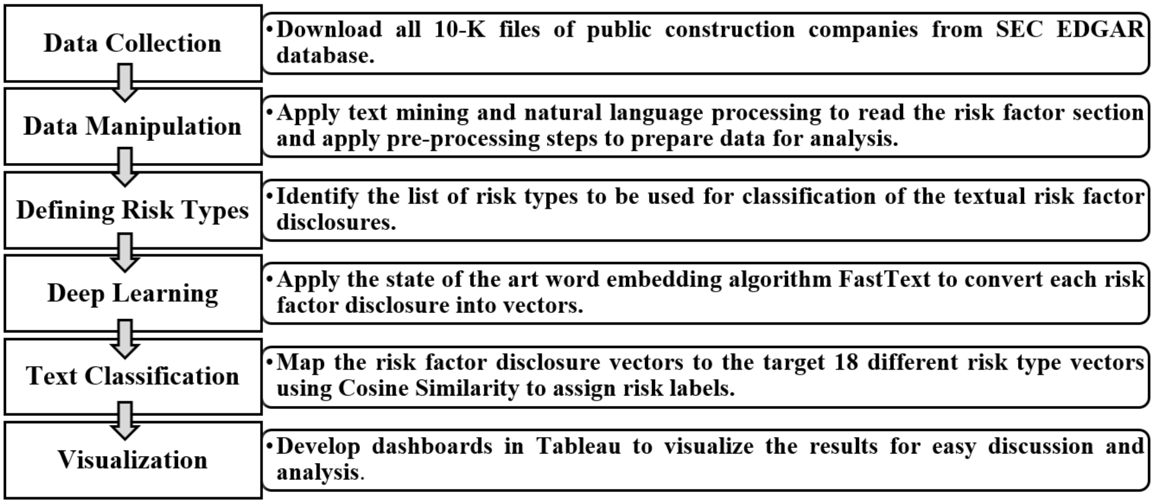

Data Collection
•Download all 10-K files of public construction companies from SEC EDGAR database.
Data Manipulation
•Apply text mining and natural language processing to read the risk factor section and apply pre-processing steps to prepare data for analysis.
Defining Risk Types
•Identify the list of risk types to be used for classification of the textual risk factor disclosures.
Deep Learning
•Apply the state of the art word embedding algorithm FastText to convert each risk factor disclosure into vectors.
Text Classification
•Map the risk factor disclosure vectors to the target 18 different risk type vectors using Cosine Similarity to assign risk labels.
Visualization
•Develop dashboards in Tableau to visualize the results for easy discussion and analysis.

Figure 2 - Flowchart of the methodology adopted for the research

CHAPTER 3. SEC 10-K FILINGS DATA

When a privately-held company wishes to grow its operations, expand its business and raise a large amount of capital for aforementioned and/or other purposes, one of the options they have is to decide to if they want to go public. The process of going public is generally undertaken by an Initial Public Offering (IPO), which enables the company to become a publicly traded and owned entity. The company is then allowed to raise money from investors all over the world, who exchange their money for a piece (or share) of the ownership of the company. The natural advantages of going public include strengthening the capital base of the company, increase in prestige and creating avenues for acquisitions and growth. However, these privileges come with increased pressure, added costs and requirement for compliance with statutory bodies, which mandates the company to impose restrictions on its management practices and trading, forces it to make its disclosures readily available to the public and dilutes the ownership and decision-making control from the private owners to the broader group of shareholders. In the United States, the federal regulatory body which enforces and ensures that all public companies are in compliance with these rules and regulations, is known as the Securities and Exchange Commission (SEC).

The SEC was established in 1930s to prevent stock manipulation and fraud, and acts as a regulatory watchdog for the United States federal government. It collects detailed documents which consist of financial and operational information of all public companies whose stock trades in the nation’s stock exchanges. The SEC ensures that the quality of the information provided by the different companies meet certain requirements. Several investors look at these filings, study the documents to understand the inner-workings of the companies and determine the health and well-being of the companies' financial future. Out of the multitude of different filings mandated by the SEC, the annual 10-K report is considered the most important and is a widely examined and scrutinized document in the financial industry.

3.1 Understanding 10-K Filings

The 10-K report ("SEC.gov | How to Read a 10-K," n.d.) is a comprehensive annual report of the company which is required to be filed with the SEC within 90 days of the end of its fiscal year. The 10-K report comprises of the several sections which include: -

a. Business: This provides an overview of the company's main operations, including its main products and services, the subsidiaries that it owns, and the different markets that it operates in.
b. Risk Factors: These outline any and all risk factors that the company faces or may face in the future. According to the SEC's website, 'some of the risk factors listed in this section may be true for the entire economy, some may apply only to the company's industry sector or geographic region, and some may be unique to the company.'
c. Selected Financial Data: This section details specific financial information about the company over the last five years. This section presents more of a near-term view of the company's recent performance.
d. Management's Discussion and Analysis of financial condition and results of operations: Also known as MD&A, this gives the company an opportunity to

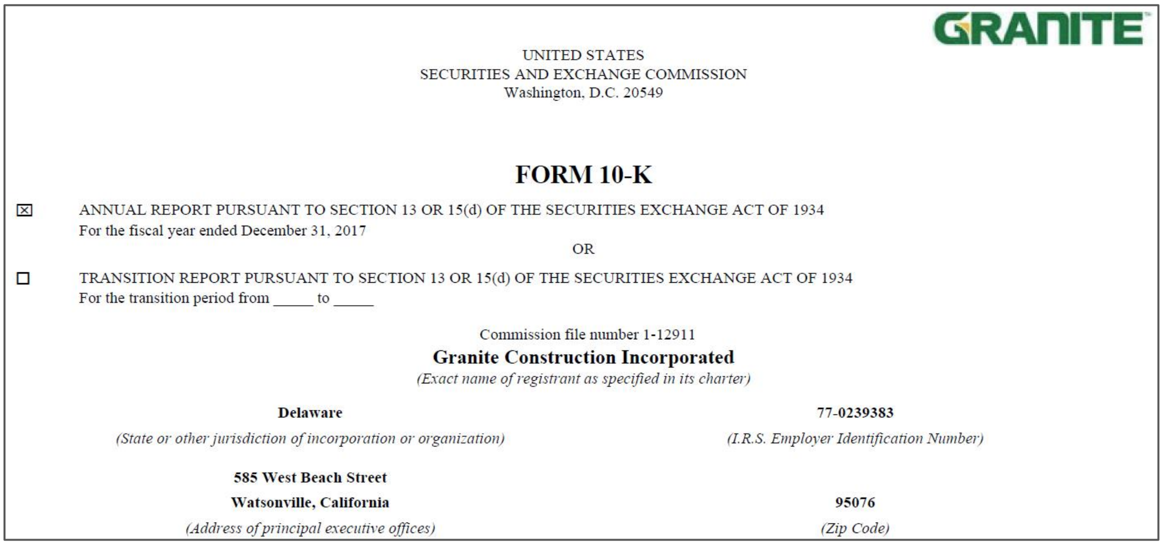
explain its business results from the previous fiscal year. This section is where the company can tell its story in its own words.

e. Financial Statements and Supplementary Data: This includes the company's audited financial statements including the income statement, balance sheets, and statement of cash flows. A letter from the company's independent auditor certifying the scope of their review is also included in this section.

The SEC staff review 10-K filings and sometimes they can provide comments to a company whose disclosures are found to be inconsistent or vague with the overall disclosure requirements. The Sarbanes Oxley Act ("Sarbanes-Oxley Act of 2002," n.d.), which is also known as the "Public Company Accounting Reform and Investor Protection Act", was enacted in July of 2002, which mandated a new set of expanded requirements for all public company boards, their managements and their accounting firms. This act requires the SEC to review all public companies' filings at least once every three years. Sometimes, the SEC staff may choose to review the financial filings of some companies much more frequently.

To provide a sense of familiarity for the subject matter inside a 10-K file, the cover page and the table of contents of the 10-K filing of Granite Construction (Granite Construction Incorporated | Form 10-K (2018) n.d.) for the fiscal year 2018 are presented as seen in Figure 3 and Figure 4. 10-K files typically have 15 schedules (or items) as seen. As per the description provided in their 10-K file for 2018, Granite Construction is one of the largest diversified heavy civil contractors and construction materials producers in the United States. They operate nationwide and serve both public and private sector clients.

Their business is organized into three reportable segments which are Construction, Large Project Construction and Construction Materials.



GRANITE

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

☒ ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the fiscal year ended December 31, 2017

OR

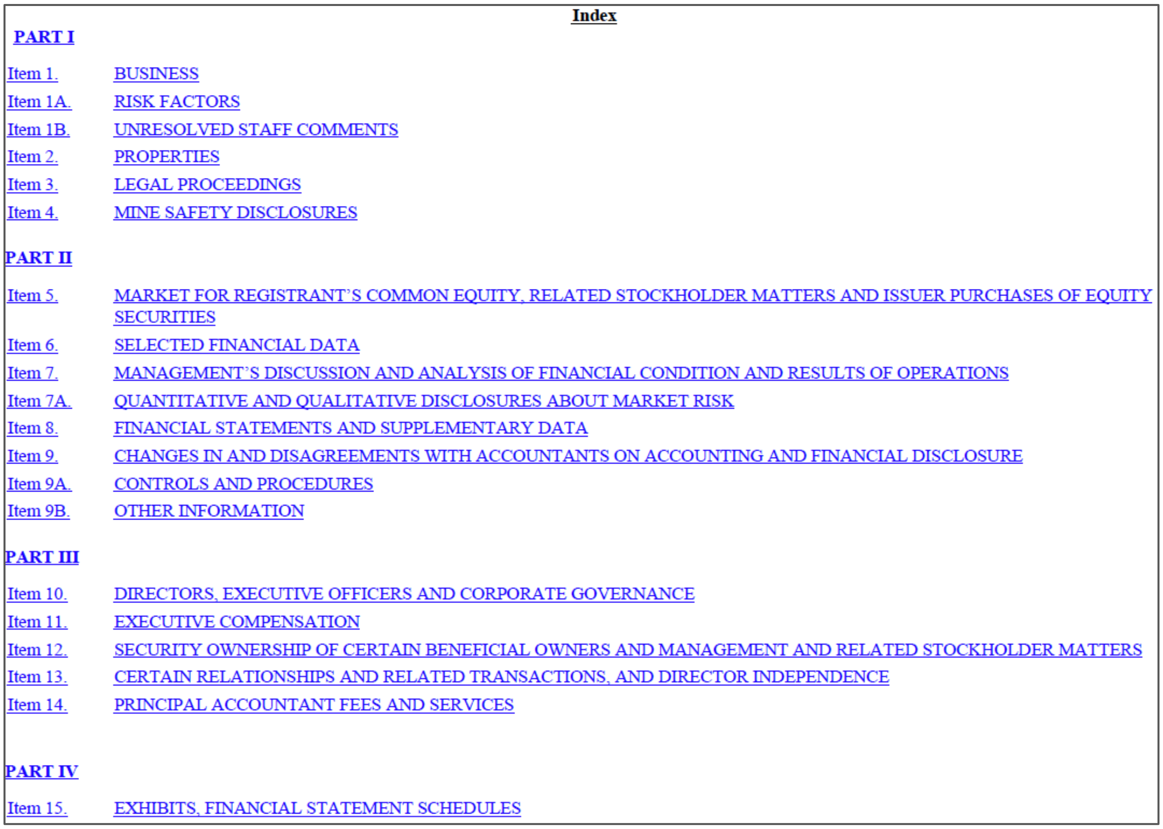
☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from _____ to _____

Commission file number 1-12911
Granite Construction Incorporated
(Exact name of registrant as specified in its charter)

Delaware	**77-0239383**
(State or other jurisdiction of incorporation or organization)	*(I.R.S. Employer Identification Number)*
585 West Beach Street **Watsonville, California**	**95076**
(Address of principal executive offices)	*(Zip Code)*

Figure 3 - Snapshot of cover page of Granite Construction's 10-K filing for 2018

The 10-K file for Granite Construction for the fiscal year 2018, was a 96-page document when downloaded in a PDF format. the file consists of rich information about the company and its financial position, complete with facts and figures.



Index

PART I	
Item 1.	BUSINESS
Item 1A.	RISK FACTORS
Item 1B.	UNRESOLVED STAFF COMMENTS
Item 2.	PROPERTIES
Item 3.	LEGAL PROCEEDINGS
Item 4.	MINE SAFETY DISCLOSURES
PART II	
Item 5.	MARKET FOR REGISTRANT'S COMMON EQUITY, RELATED STOCKHOLDER MATTERS AND ISSUER PURCHASES OF EQUITY SECURITIES
Item 6.	SELECTED FINANCIAL DATA
Item 7.	MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS
Item 7A.	QUANTITATIVE AND QUALITATIVE DISCLOSURES ABOUT MARKET RISK
Item 8.	FINANCIAL STATEMENTS AND SUPPLEMENTARY DATA
Item 9.	CHANGES IN AND DISAGREEMENTS WITH ACCOUNTANTS ON ACCOUNTING AND FINANCIAL DISCLOSURE
Item 9A.	CONTROLS AND PROCEDURES
Item 9B.	OTHER INFORMATION
PART III	
Item 10.	DIRECTORS, EXECUTIVE OFFICERS AND CORPORATE GOVERNANCE
Item 11.	EXECUTIVE COMPENSATION
Item 12.	SECURITY OWNERSHIP OF CERTAIN BENEFICIAL OWNERS AND MANAGEMENT AND RELATED STOCKHOLDER MATTERS
Item 13.	CERTAIN RELATIONSHIPS AND RELATED TRANSACTIONS, AND DIRECTOR INDEPENDENCE
Item 14.	PRINCIPAL ACCOUNTANT FEES AND SERVICES
PART IV	
Item 15.	EXHIBITS, FINANCIAL STATEMENT SCHEDULES

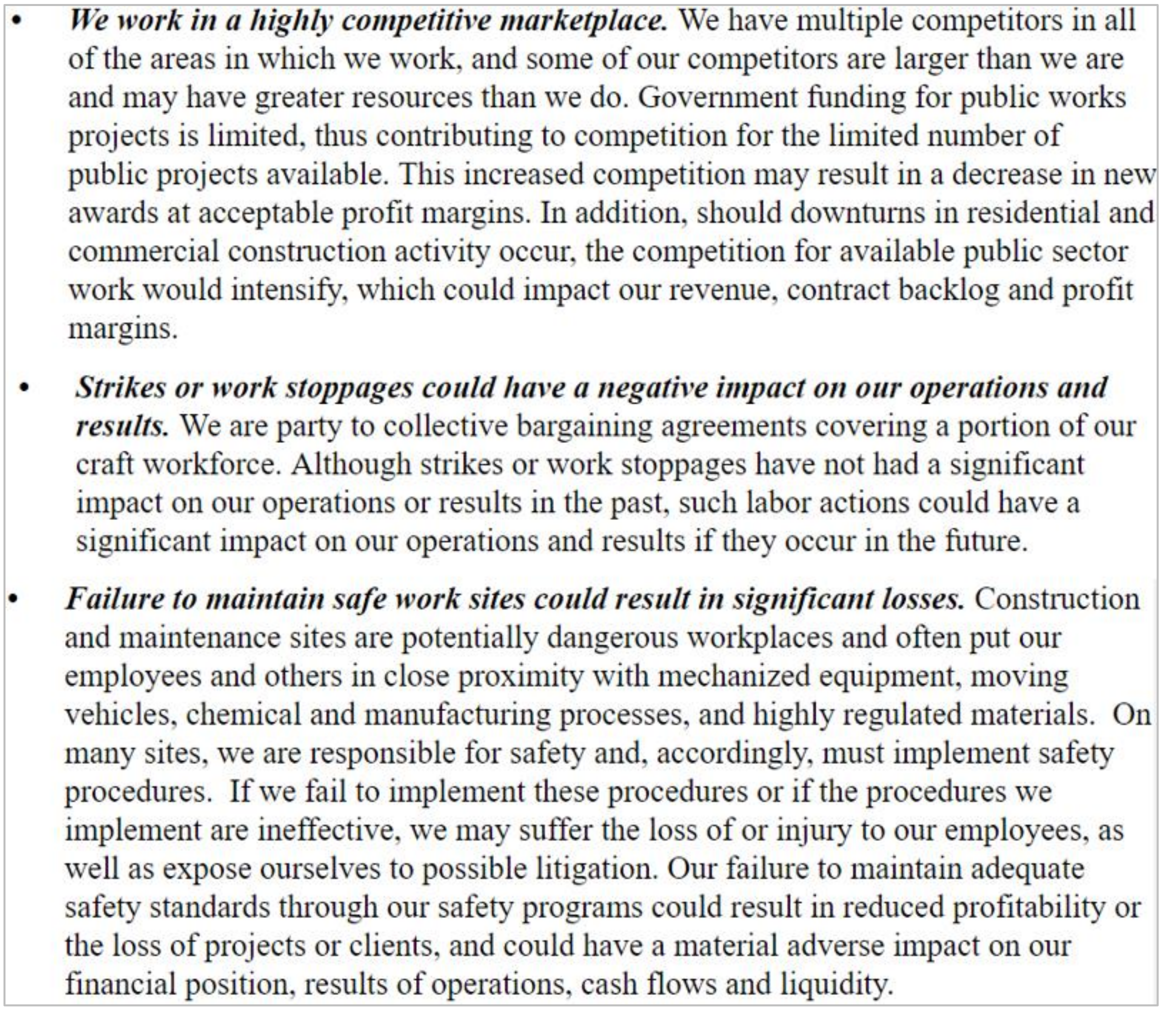
Figure 4 - Snapshot of Table of Contents of Granite Construction's 10-K filing for 2018

3.2 Understanding 'Item 1A - Risk Factors'

In 2005, the SEC began requiring an 'Item 1A - Risk Factors' to be a separate section of disclosure in the 10-K reports (SEC.gov | How to Read a 10-K n.d.). In this section the company discloses the most significant risks that could harm its business. It can range from risks inherent to a specific industry to something that is broader. A typical 'Item 1A - Risk Factors' section of a company lists various risk factors in free-flowing text. The guidelines provided by the SEC state that the risk factors section must be written using plain English principles that include short sentences, definite, concrete and everyday words, active voice, bullets wherever possible, no legal or highly technical jargon and no multiple negatives.

The goal is to provide any potential investor with clear information and full disclosure of risk factors affecting the company.

Figure 5 provides a glimpse of three risk factors presented in the 10-K filing of Granite Construction for the fiscal year 2018. Typically, the risk factors section consists of a bullet point header which is usually in bold and italics font, that captures the main essence of the risk factor, which is followed by a short paragraph describing the risk factor in some more detail. The first risk factor shown in the image details risks due to the company operating in a competitive market place as there are other construction companies having higher revenues and resources, that may be a challenge for Granite Construction in being able to bid and win the award of new public projects. The second risk factor disclosure deals with possibility of delays due to work stoppages and labor strikes which can in turn lead to a negative impact on the day-to-day operations of the company and its financial condition. It details that the company is under a collective bargaining agreement with a segment of its workforce and any potential disagreements with the labor unions can be a major issue. Finally, the third risk disclosure details risks related to safety in the work site. It illustrates that construction sites are inherently dangerous workplaces as their workers and employees have to work in close proximity of mechanized equipment, vehicles, chemicals and other hazardous material. In such an environment, the risk of potential injury/fatality exposes the company to litigation and can also lead to reduce in profitability and a negative impact on the overall financial position of the company.



- ***We work in a highly competitive marketplace.*** We have multiple competitors in all of the areas in which we work, and some of our competitors are larger than we are and may have greater resources than we do. Government funding for public works projects is limited, thus contributing to competition for the limited number of public projects available. This increased competition may result in a decrease in new awards at acceptable profit margins. In addition, should downturns in residential and commercial construction activity occur, the competition for available public sector work would intensify, which could impact our revenue, contract backlog and profit margins.
- ***Strikes or work stoppages could have a negative impact on our operations and results.*** We are party to collective bargaining agreements covering a portion of our craft workforce. Although strikes or work stoppages have not had a significant impact on our operations or results in the past, such labor actions could have a significant impact on our operations and results if they occur in the future.
- ***Failure to maintain safe work sites could result in significant losses.*** Construction and maintenance sites are potentially dangerous workplaces and often put our employees and others in close proximity with mechanized equipment, moving vehicles, chemical and manufacturing processes, and highly regulated materials. On many sites, we are responsible for safety and, accordingly, must implement safety procedures. If we fail to implement these procedures or if the procedures we implement are ineffective, we may suffer the loss of or injury to our employees, as well as expose ourselves to possible litigation. Our failure to maintain adequate safety standards through our safety programs could result in reduced profitability or the loss of projects or clients, and could have a material adverse impact on our financial position, results of operations, cash flows and liquidity.

Figure 5 - Snapshots of some risk factors in the 10-K file of Granite Construction for the fiscal year 2018

3.3 Data Collection: Identifying all Public Construction Firms in the SEC database

The SEC EDGAR database consists of an online directory of all the public companies with their financial filings. It conveniently indexes the extensive list of companies in its records by various different types of variables such as filing date, Central Index Key (CIK), filing type, name of the company, and link to the files and company location. For the purposes of this research, the ‘filing type’ filter is restricted to ‘10-K’, the location is restricted to include only the companies based in the United States. It was in

2005 when the SEC began requiring an ‘Item 1A - Risk Factors’ to be a separate section of disclosure in the 10-K reports. Hence for the current research, the data was first collected starting January 2006 until August 2019.

In addition to these key variables, one of the essential variables for filtering the dataset to obtain the companies belonging to the area of interest for this specific project, is the Standard Industrial Classification (SIC) code of the companies. The SIC classification is a system for classifying industries by a four-digit code (Division of Corporation Finance SIC Code List n.d.). The first 3 digits of the SIC code indicate the industry group, and the first two digits indicate the major group. Out of the 12 broad divisional groups, one is Construction. The construction divisional group has several sub-groups based on the type of services provided by the company. The SEC EDGAR database has 8 sub-industry SIC codes that belongs to the construction industry as shown in Table 4 with their respective brief description and examples.

Table 4 - SIC Codes related to Construction Industry

SIC code and Classification name
1520 (General Building Contractors - Residential Buildings)
General contractors primarily engaged in construction (including new work, additions, alterations, remodeling, and repair) of residential buildings. *Ex: Lennar Corp, Brookfield Homes Corp*
1531 (Operative Builders)
Builders primarily engaged in the construction of single-family houses and other buildings for sale on their own account rather than as contractors. *Ex: D.R Horton, MDC Holdings, NVR inc.*

Table 4 Continued

1540 (General Building Contractors - Nonresidential Buildings)

General contractors primarily engaged in the construction (including new work, additions, alterations, remodeling, and repair) of non-residential buildings.

Ex: Tutor Perini, Sports Field Holdings

1600 (Heavy Construction other than Building Construction - Contractors)

This group includes general contractors primarily engaged in heavy construction other than building, such as highways and streets, bridges, sewers, railroads, irrigation projects, flood control projects and marine construction.

Ex: Jacobs Engineering, Fluor Corp, Sterling Construction

1623 (Water, Sewer, Pipeline, Comm. and Power Line Construction)

General and special trade contractors primarily engaged in the construction of water and sewer mains, pipelines, and communications and power lines.

Ex: Aegion Corp, DYCOM industries, Mastec inc.

1700 (Construction Special Trade Contractors)

This group includes special trade contractors who undertake activities which include painting, electrical work, carpentry work, plumbing, heating, air-conditioning, roofing, and sheet metal work.

Ex: Matrix Service, Layne Christensen Company

3531 (Construction Machinery & Equipment)

Establishments primarily engaged in manufacturing heavy machinery and equipment of a type used primarily by the construction industries, such as bulldozers, concrete mixers, cranes, dredging machinery, pavers, and power shovels.

Ex: Caterpillar, Astec, Gencor

8711 (Engineering Services)

Establishments primarily engaged in providing professional engineering services. Establishments primarily providing and supervising their own engineering staff on temporary contract to other firms are included in this industry.

Ex: AECOM, CH2M, Hill International

The SEC does not assign codes, however, when a company registers its IPO, it selects a SIC code based on its primary source of revenue. If the primary source of revenue for a company changes, the company can reach out to the SEC to alter the SIC code they're using.

3.3.1 Downloading a master index of all SEC Filings from 'Python-Edgar'

The SEC EDGAR maintains a rich database of all financial filings since the year 1993 ("Directory listing of full-index/," n.d.). The database is provided at a year level which is further dis-aggregated at a quarter level. For this research, the open-source programming language Python 3 was used as it provides extensive open-source resources and libraries for implementation of data science and text mining methods. For the first step, Python library named 'Python - Edgar' ("Python-edgar · PyPI," n.d.) is used to download the master index database with ease. This database contains the '*CIK code', 'name of company', 'filing type', 'filing date', 'text file web-link'* and the *'html file web-link'*. In total, since 1993, until the time of final iteration of data collection in August 2019, the total number of records were found to be 299,639 total 10-K SEC filings. These consists of all types of public companies and their 10-K financial filings.

3.3.2 Identifying the list of construction companies and their CIK codes

Since the database downloaded from 'Python-Edgar' consisted of the name of the company and the link to download its 10-K filing by the CIK code, the next step was to identify the CIK codes for all the construction industry public companies. In order to do that, the SEC EDGAR website provides a very flexible search page ("Comprehensive Search Page," n.d.) which lets the user search for the companies and their associated CIK codes by using the SIC classification. As earlier listed in Table 4, there are eight (8) sub-industry classifications within the construction domain that fit these criteria. The search results of these eight SIC codes were collected in an Excel file that contained '*CIK code'*, *'name of company'*, *'location'* and *'SIC Code'*. These CIK codes of the requisite construction companies are merged with the database of web-links of all 10-K filings to obtain the requisite table containing '*CIK code'*, *'name of company'* and *'html file web-link'*. Since the scope of the research project is limited to construction companies located within the United States and the time-period for the risk factors data begins from 2006, these filters are applied to finally arrive at 1,166 rows. Each row is essentially a *company name* and its *10K file web-link* data.

3.3.3 Downloading the 1166 10-K files in HTML format

At this stage, the web-links for the total database of 1166 10-K files for all US construction industry firms since 2006 until August 2019, is available. Next step was to download all these 1166 files for analysis. To achieve this, several Python libraries typically used for web-scraping, such as 'requests', 're', 'beautiful-soup' (Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation n.d.; re — Regular Expression operations — Python 3.8.0 documentation n.d.; requests · PyPI, n.d.) were utilized. It is to be noted that the web-links obtained from the 'Python-Edgar' library routes the user to the index page of the 10-K file, and not the 10-K file itself. The index page consists of the link to the actual 10-K file; hence a Python program is written to find the exact extension for the file link. Once the exact web-link for the 10-K file in html format is retrieved, the file is downloaded using code. This process is programmatically automated for all the 1166 files. Ultimately, the files are downloaded in html format to facilitate analysis.

3.3.4 Extracting 'Item 1A - Risk Factors' section from all the 10-K files

Once the entire database of 1166 10-K files were downloaded in html format, the next step was to programmatically extract the section of focus, i.e. the 'Item 1A- Risk Factors' section. This process was found to be particularly challenging as there did not exist a clear-cut easy pattern among all the files, such that a simple piece of code would facilitate this extraction. The author had to carefully build a heuristics-based algorithm after multiple iterations of trial and error, which accounted for the html structure of the downloaded files. Primarily, two types of approaches were implemented which include (a) programmatically searching for the text 'risk factors' including all combinations of the phrase in uppercase, lowercase, with all combinations of special characters in between etc. (b) finding the hyperlink of risk factors section if it is present in the contents table. Once the location on the risk factors section was found within the file, the code was crafted to carefully find the end position of the section as well, on similar lines of finding its starting location in the 10-K file. Having identified the exact start and end points for the risk factors section, all the text between these two locations was extracted by code.

An important point of information at this point is that not all the downloaded 10-K files have a clear Item 1A - Risk Factor section. Especially some of the 10-K files between the period 2006-2009 do not have it at all, back when this was still a relatively new requirement by the SEC. About 15% of the total downloaded files were found to be unfit for further analysis. The reasons to discard them included them either not having a clearly defined risk factors section or only a small and vague risk description (5 risk factor disclosures or less). Finally, after filtering out the ill-fitting 10-K files, the universe of 10-K files ultimately used in the analysis was a total of 995 10-K files as shown in Table 5.

Table 5 - Publicly traded construction companies by SIC code and their total number of 10-K files for the time-period January 2006 - August 2019

SIC code and Classification name	# of Companies	Total # of 10-Ks
1520 (General Building Contractors - Residential Buildings)	13	59
1531 (Operative Builders)	32	259
1540 (General Building Contractors - Nonresidential Buildings)	9	47
1600 (Heavy Construction other than Building Construction - Contractors)	11	108
1623 (Water, Sewer, Pipeline, Comm. and Power Line Construction)	9	103
1700 (Construction Special Trade Contractors)	29	147
3531 (Construction Machinery & Equipment)	10	79
8711 (Engineering Services)	24	193
Total	**137**	**995**

It was found that the final universe of 995 10-K files comprised of a total of 137 unique construction companies (an average of 7-8 10-K files per company over a 14-year period). A complete list of all the companies whose 10-K files were used is provided in

Appendix A. It was found that some of these companies have filing information for all the years since 2006, whereas other companies had data for only a few year(s). This is dependent on various factors which include: -

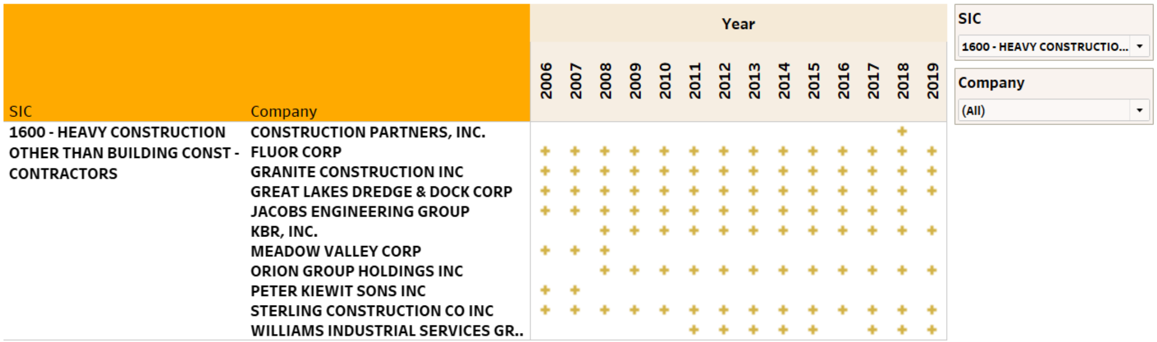
- Year of inception: - a new company which has been set up after a certain year (for ex: 2015), will not have any 10-K filings before its inception.
- Year of going public: - availability of filings depend on when the company decided to go public.
- Size of the company: - Smaller reporting companies ("SEC.gov | Smaller Reporting Companies," n.d.) (generally with a comparatively smaller overall revenue etc.) do not have same stringent filing requirements as large companies.
- Events like bankruptcy, private buyouts etc.: - these and other similar scenarios are possible reasons why a company may discontinue filing 10-Ks.
- Negligence and incorrect filings: - even after SEC imposing strict requirements, some of the companies fail to meet the requirements and therefore their data may be incomplete or absent.

Table 6 provides a year-wise distribution of all the downloaded 10-K files. It is seen that on a broad level, there were a total of 60-80 10-K files every year. Since the final iteration of data collection was done in August 2019, the number of filings for the year 2019 is lower than the expected number for the entire year.

Table 6 - Year-wise distribution of 10-K files for the filtered dataset

Year	Total # of 10-Ks
2006	61
2007	64
2008	69
2009	76
2010	73
2011	74
2012	74
2013	73
2014	81
2015	81
2016	73
2017	72
2018	68
2019 (until August)	56
Total	**995**

For the purposes of demonstration of the methodology, results and discussion, from this point onwards, SIC 1600 - 'Heavy Construction other than Building Construction - Contractors' is used as an example. The research makes use of the popular data visualization tool Tableau as the front-end interface. Figure 6, generated in Tableau, demonstrates all the companies that fall within the SIC 1600 group, with the distribution of the files collected for them over the years (until August 2019).



		Year													
SIC	Company	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
1600 - HEAVY CONSTRUCTION OTHER THAN BUILDING CONST - CONTRACTORS	CONSTRUCTION PARTNERS, INC.													+	
	FLUOR CORP	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	GRANITE CONSTRUCTION INC	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	GREAT LAKES DREDGE & DOCK CORP	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	JACOBS ENGINEERING GROUP	+	+	+	+	+	+	+	+	+	+	+	+	+	
	KBR, INC.			+	+	+	+	+	+	+	+	+	+	+	+
	MEADOW VALLEY CORP	+	+	+											
	ORION GROUP HOLDINGS INC			+	+	+	+	+	+	+	+	+	+	+	+
	PETER KIEWIT SONS INC	+	+												
	STERLING CONSTRUCTION CO INC	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	WILLIAMS INDUSTRIAL SERVICES GR..						+	+	+	+	+		+	+	+

SIC
1600 - HEAVY CONSTRUCTIO...

Company
(All)

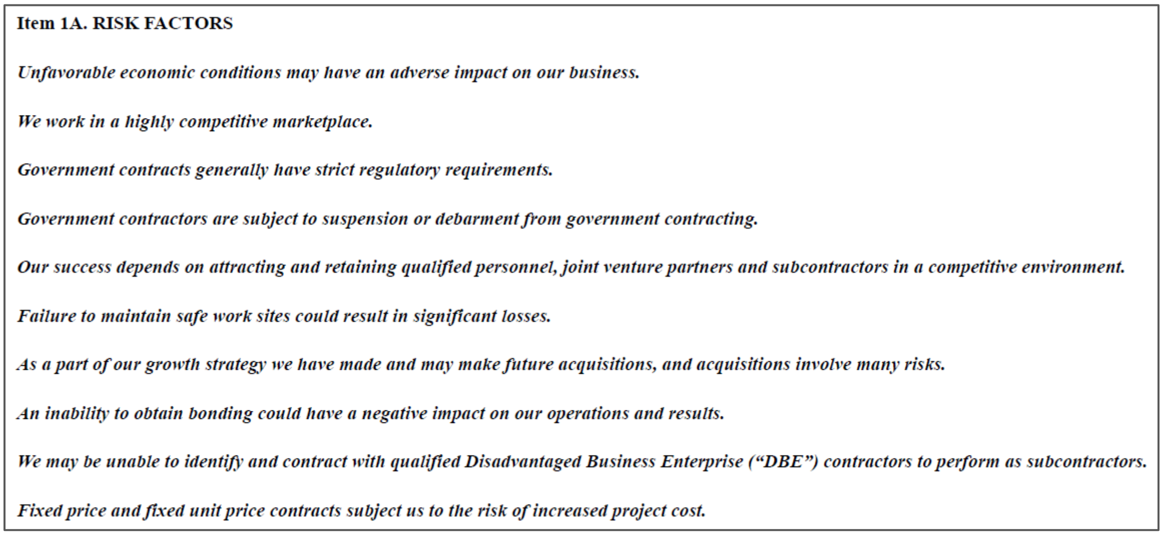
Figure 6 - Public Companies of SIC 1600 and their 10-K filings collected by year (data until August 2019)

3.4 Data Manipulation

3.4.1 Extracting the risk factor disclosures from 'Item 1A - Risk Factors' section of the 10-K dataset

Once the Item 1A - Risk Factors section is successfully extracted from the 10-K files, the next step is to extract the main sentences in bold/italic font which serves as the header for each risk factor disclosure. As seen earlier in the example presented in Figure 5, the risk factor section consists of a main header sentence that captures the main essence of the risk factor disclosure, followed by a paragraph describing it. For the purposes of this study, as the focus is on using a text mining-based technique to automatically identify and classify the different risk factors in the files, only the main headings of each risk factor disclosure are extracted in order to get rid of confounding information and only retain useful text. Figure 7 presents an example to illustrate a sample of the extracted risk factor disclosures obtained from the Item 1A- Risk Factors section of the 10-K file of Granite

Construction for the year 2018. Notice that only the text in bold and italics is retained and the paragraphs of text describing each risk factor disclosure is filtered out.



Item 1A. RISK FACTORS

Unfavorable economic conditions may have an adverse impact on our business.

We work in a highly competitive marketplace.

Government contracts generally have strict regulatory requirements.

Government contractors are subject to suspension or debarment from government contracting.

Our success depends on attracting and retaining qualified personnel, joint venture partners and subcontractors in a competitive environment.

Failure to maintain safe work sites could result in significant losses.

As a part of our growth strategy we have made and may make future acquisitions, and acquisitions involve many risks.

An inability to obtain bonding could have a negative impact on our operations and results.

We may be unable to identify and contract with qualified Disadvantaged Business Enterprise ("DBE") contractors to perform as subcontractors.

Fixed price and fixed unit price contracts subject us to the risk of increased project cost.

Figure 7 - A sample of programatically extracted Risk Factor Disclosures for Granite Construction for its 10-K filing of 2018.

The process of extracting these risk factor disclosures in Python for each file is achieved by developing a rule-based methodology using Beautiful Soup and Regular Expressions open-source libraries. The entire text of the risk factors section html file was loaded in Python and then the html tag markers for bold and italics tags were located. Search patterns were coded using Regular Expression Python library to narrow down on the important and required pieces of text and extracted in string format. Care was taken to make sure that each extracted individual risk factor disclosure was correctly matched with the *company name* and *year* in which it was filed for disclosure. A total of 29,398 risk factor disclosures were extracted from 995 files (an average of about 30 risk factor disclosure per file over the entire time period). These 29,398 risk factor disclosures were base unit of analysis for the research project.

3.4.2 Pre - Processing and Text Cleaning

Any implementation of natural language processing or text mining requires a series of pre-processing and text cleaning steps to obtain text data that can be further used for data science and machine learning analysis. The extracted risk factor disclosures were subjected to a series of steps which are as follows:

- Convert all the data into lowercase to ensure consistency.
- Strip out all unwanted html tags and symbols like ‘\\xa0’, ‘\\t’, ‘<div>’ etc.
- Remove all punctuation, digits and symbols like @, $, *, ^, % etc.
- Remove all proper nouns.
- Remove any words that are less than equal to a length of 2 characters.
- Remove a host of stop-words like ‘and’, ‘or’, ‘not’, ‘although’, ‘but’ etc., that do not contribute to the determination of risk classification.
- Remove words which are too commonly found and do not contribute to risk classification.
- Stemming and Lemmatization (“Stemming and lemmatization,” n.d.) to retain the root form of words.

The pre-processing and text cleaning steps are intended to retain meaningful data and eliminate a significant portion of confounding information.

3.4.3 Correction for mis-spellings in the textual data

Even though the 10-K files are professionally filed and audited by public companies, there were some instances of mis-spellings found in the data. To correct any mis-spelled words and to remove any rogue text that had remained after the text extraction and cleaning process, a database of all possible words in the English language was created using WordNet (“WordNet | A Lexical Database for English,” n.d.) and Spacy Word Vectors (“English · spaCy Models Documentation,” n.d.). Any word in the corpus that didn’t belong in the WordNet and Spacy word list was examined to check for mis-spellings and to see if they were just gibberish. The mis-spellings were handled for further analysis. Table 7 presents some examples some risk factor disclosure sentences after pre-processing, text cleaning and handling of mis-spellings to obtain the final form which retains only the signal and gets rid of noise.

Table 7 - Pre-processing and Text Cleaning applied to a sample of Risk factor disclosures of Granite Construction for its 10-K filing of 2018

Full risk disclosure sentences	After text cleaning
Unfavorable economic conditions may have an adverse impact on our business.	unfavorable economic condition adverse impact business
We work in a highly competitive marketplace.	work highly competitive marketplace
Government contracts generally have strict regulatory requirements.	government contract generally strict regulatory requirement
Government contractors are subject to suspension or debarment from government contracting.	government contractor subject suspension debarment government contract
Our success depends on attracting and retaining qualified personnel, joint venture partners and subcontractors in a competitive environment.	success depend attract retain qualified personnel joint venture partner subcontractor competitive environment

Table 7 Continued

Failure to maintain safe work sites could result in significant losses.	failure maintain safe work site result significant loss
As a part of our growth strategy we have made and may make future acquisitions, and acquisitions involve many risks.	part growth strategy make make future acquisition acquisition involve many
An inability to obtain bonding could have a negative impact on our operations and results.	inability obtain bonding negative impact operation result
We may be unable to identify and contract with qualified Disadvantaged Business Enterprise (DBE) contractors to perform as subcontractors.	unable identify contract qualified disadvantaged business enterprise dbe contractor perform subcontractor
Fixed price and fixed unit price contracts subject us to the risk of increased project cost.	fix price fix unit price contract subject increase project cost

CHAPTER 4. IDENTFYING AND CLASSIFYING RISK TYPES

4.1 Defining Risk Types

The risk factor disclosure requirements by the SEC allows the different publicly traded companies to list various risk factors in their own language using plain English words. The flexibility provided by the SEC is desirable to ensure room for the companies to be able to express issues concerning them in the best way possible. However, for the present research, from the perspective of classifying the risk factors into easily identifiable and interpretable buckets, the main issue that is faced is that there is no standard classification or categories that need to be adhered to by all firms in disclosing their risk factors. The risk disclosures can discuss very specific risks to the company or they can be about general issues. Hence, a gap of developing an appropriate classification of risk types that can be used to track different areas of risks across different firms and over time in the construction industry, is encountered.

To achieve this objective, a content analysis approach was used to classify risk disclosures into several generalizable risk categories. Special consideration was placed to keep the total risk types to a reasonable number for straightforward interpretation as this research approaches the risk identification problem from a broader industry level. First, a random sample of 500 risk factor disclosures (out of the database of 29,398 risk factor disclosures) was selected. Next, the content of these risk disclosures was carefully analyzed and placed into 18 distinct risk types, which were found to be salient and representative of the underlying risk data. Table 8 identifies the 18 distinct risk types, their description and several keywords that assist in the classification. The identified keywords are important terms that trigger the decision of risk type classification. These keywords were manually identified during the content analysis of the sample of 500 risk factor disclosures. While conducting this exercise, the list of 18 risk types was reached after examining first 100 risk factor disclosures itself. For the remaining 400 risk factor disclosures, the author did not find a need to define a new risk category as all of the risk disclosures were attributable to the 18 risk types identified.

Table 8 - Identified risk types with their description and sample of important keywords

Risk Type	Description of Risk Type	Sample of Important Keywords
Accounting and Financial Reporting Risks	Risk related to any accounting and reporting issues and changes in the regulatory requirements.	accounting, accounting policy, financial reporting, financial statement, and percentage completion method
Business Risks - Operational and Financial	Risk related to the business operations, growth or financing of business processes.	acquisition, business, cash flow, financial condition, growth, illiquid, and operating result
Competition Risks	Risk related to competition from other players in the industry.	compete, competition, competitive, and competitor
Cost Risks	Risk related to costs incurred, material price, and unforeseen expenses.	cost, cost increase, incur, and material price
Financial Market Risks	Risk related to being a public firm whose stock trades in financial markets.	common stock, dividend, equity, market price, and share price
Governmental, Contractual and Regulatory Risks	Risk related to changes in governmental policies, adherence with contracts and increasing regulation.	contract government, federal, government contract, and regulation

Table 8 Continued

Human Resource Risks	Risk related to the human resources including management and its impact in the business.	attract, board member, employee, key personnel, and recruit
Land, Property and Inventory Risks	Risk related to tangible assets such as land, property and inventory.	inventory, land, and land inventory
Legal, Claims, Liabilities and Dispute Risks	Risk related to all legal dealings of the company, claims, disputes and any existing or future litigation.	claim, contract dispute, indemnification, insurance, legal, legislation, and liability
Macro Risks	Risk related to the economy, demand and supply of goods and services, and international factors.	cyclical, demand, downturn, economic, geopolitical, inflation, natural gas, and oil
Natural and Manmade Disasters Risks	Risk related to disasters, both natural and man-made.	terrorism, war, earthquake, flood, force majeure, natural disaster, and weather
Reputational and Intangible Risks	Risk related to intangible assets and harm to the reputation of the company.	goodwill, intangible asset, intellectual property, patent, and reputation
Safety Risks	Risk related to safety in the worksite.	dangerous, death, injury, occupational safety health administration, and security
Tax Risks	Risk related to tax policies and tax regulation.	income tax, internal revenue service, tax law, and tax rate
Technology Risks	Risk related to technology, cyber security and automation.	breach, computer, cybersecurity, information technology, and software
Third Party Risks	Risk related to dealings with third parties, sub-contractors, and joint venture.	counterparty, joint venture, subcontractor, and third party

Table 8 Continued

Time, Delay and Uncertainty Risks	Risk related to dynamic nature of the business, and uncertainty in time of completion and delays.	change order, delay, late completion, timely, and work stoppage
Work Quality and Error Risks	Risk related to quality of the work produced and errors in the product/service.	assess quality, construction defect, design error, and quality control

These 18 risk types are not mutually exclusive. The focus of the 10-K risk factor disclosures is on describing the risk from the finance and accounting perspective. A large percentage of the risk factors discuss the business risks facing the company with specific attention on the financial position and the operations of the company. A significant number of other risk factors discuss the impact of risk from an operational and financial perspective. Hence, by the nature of the risk factor disclosures, the 'Business Risks - Operational and Financial' is expected to be a prevalent risk in most cases. In addition to the identification of the most frequently discussed risk types, special attention was directed to explore the risk types that are less frequent, but could be considered critical to the business bottom line of firms in the construction industry. Each risk factor disclosure is allowed to be classified as none, one or more than one risk type. By design, the classification includes risk types that can be considered as the main focus for a risk factor disclosure as well as other risk types that are affected as a result. To demonstrate the different risk types identified, Table 9 presents some examples of risk factor disclosures for Granite Construction from its 10-K file for the year 2018, and the risk types that they are manually classified into.

Table 9 - Some risk factor disclosures and their manual risk type classification from the 10-K file of Granite Construction of 2018

Risk factor disclosures	Risk type classification
Accounting for our revenues and costs involves significant estimates.	Accounting and Financial Reporting Risks
An inability to obtain bonding could have a negative impact on our operations and results.	Business Risks - Operational and Financial
We work in a highly competitive marketplace.	Competition Risks
Fixed price and fixed unit price contracts subject us to the risk of increased project cost.	Cost Risks
Rising inflation and/or interest rates could have an adverse effect on our business, financial condition and results of operations.	Macro Risks; Business Risks - Operational and Financial
Force majeure events, including natural disasters and terrorists' actions, could negatively impact our business, which may affect our financial condition, results of operations or cash flows.	Natural and Manmade Disasters Risks; Business Risks - Operational and Financial
Design-build contracts subject us to the risk of design errors and omissions.	Work Quality and Error Risks
Failure to maintain safe work sites could result in significant losses.	Safety Risks

Our success depends on attracting and retaining qualified personnel, joint venture partners and subcontractors in a competitive environment.	Human Resource Risks;
	Third Party Risks;
	Competition Risks

Table 9 Continued

A change in tax laws or regulations of any federal, state or international jurisdiction in which we operate could increase our tax burden and otherwise adversely affect our financial position, results of operations, cash flows and liquidity.	Tax Risks;
	Governmental, Contractual and Regulatory Risks;
	Legal, Claims, Liabilities and Dispute Risks

Changes to our outsourced software or infrastructure vendors as well as any sudden loss, breach of security, disruption or unexpected data or vendor loss associated with our information technology systems could have a material adverse effect on our business.	Technology Risks;
	Third Party Risks;
	Business Risks - Operational and Financial

Strikes or work stoppages could have a negative impact on our operations and results.	Time, Delay and Uncertainty Risks

4.2 Topic Modeling using Latent Dirichlet Allocation (LDA)

The overarching goal of the research project is to synthesize a methodology to apply automated text classification to unstructured textual risk disclosures. In order to do that, the specific objective is to develop a model that can parse through the textual risk factor disclosures, identify the mentions of various risk types and be able to classify them into the 18 risk type classes that are of interest. To achieve this, it is important to process and transform noisy, unstructured text data into a structured and vectorized numeric format that the computer is able to interpret. The aim here is to capture as much information as possible that is contained in the text, and transfer it into a numeric format, which facilitates classification later. Representing text in form of numbers to conduct any kind of analysis is a challenge task that is an important subject of research in natural language processing. Depending on how this transformation is performed, the output and results of such models and algorithms can be significantly different. Traditionally, most of the text mining methods have been derived from some variation of a count-based method applied to the frequency of words that exist in the text. The most popular count-based method is known as Bag of Words model. This includes techniques, such as term frequency, Tf-Idf (term frequency-inverse document frequency) and n-grams of words (Rajaraman and Ullman 2011). Topic modelling methods, such as Latent Dirichlet Allocation (LDA) (Blei et al. 2003) and Latent Semantic Analysis (LSA) (Dumais 2005), are also an extension of these count-based methods.

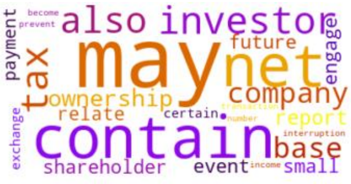
The author had successfully implemented the LDA algorithm on a research project ‘Application of Text Mining to Identify Patterns in Construction Defect Litigation Cases’ in which the technique was used to identify top fourteen (14) topics (or themes) emerging from a dataset of 59 legal case documents. The research project which is summarized in Appendix B, was an example of a successful collaboration with construction law experts to apply the up and coming field of text mining into the niche of construction industry legal research. The author spent a good portion of his 2nd and 3rd years of doctoral study working on this project and it served as a great learning experience for him and also paved the way for his ultimate doctoral research proposal and dissertation. Having applied LDA on a past research project, the author naturally decided to first try implementing it again to tackle the problem of classifying unstructured textual risk disclosures into the defined risk types.

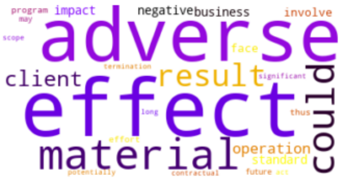
Topic modelling is essentially a sub-field in unsupervised text mining that provides the capability to organize, understand and summarize large collections of textual information. It can be used to discover hidden topical patterns present in the corpus, annotate the documents according to these topics, and use these annotations to organize, search and summarize texts. Overall, topic modelling is described as a method for finding a group of words (i.e. topic) from a collection of documents that best represent the information in the collection. LDA is a commonly used algorithm in topic modelling for natural language processing applications. In LDA, each document is viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it.

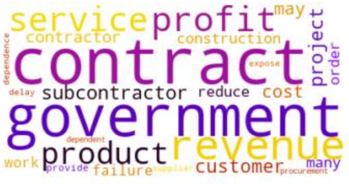
4.2.1 Implementation of LDA in Python

The actual implementation of the LDA algorithm was done in Python using libraries such as Natural Language Toolkit (NLTK) ("Natural Language Toolkit — NLTK 3.4.5 documentation," n.d.) and Gensim (Gensim: Topic modelling for humans n.d.) for developing the model and pyLDAvis (Sievert and Shirley 2014) for visualizing the results. All the risk factor disclosures obtained after the data extraction and data manipulation steps described in CHAPTER 3, served as the input data for the LDA model. The number of topics, which is a user-defined parameter for LDA, was set to 18. Table 10 shows the word clouds representing the distribution of the words associated with each of the 18 topics obtained after running LDA analysis on the entire dataset of risk factor disclosures. Each topic is a collection of words that closely capture the underlying theme presented by that topic. The topics are visualized using word clouds, and the size of each word is proportional to the weight or the probability of that word occurring in the topic.

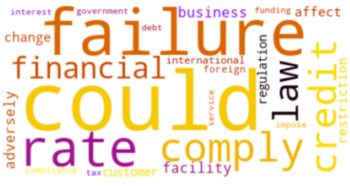
Table 10 - Word Clouds depicting Topic Word Distribution of the 18 Topics obtained after LDA analysis on the risk disclosure database

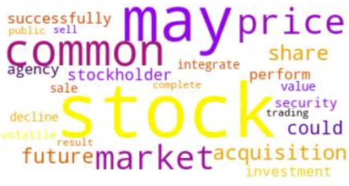


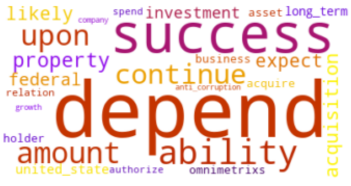




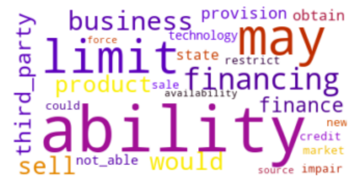
Topic #1 **Topic #2** **Topic #3**

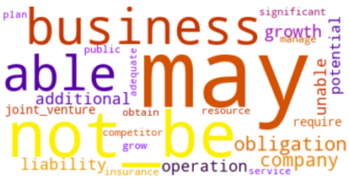


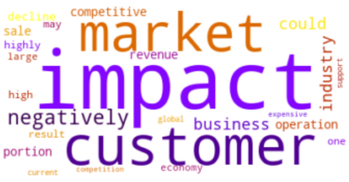




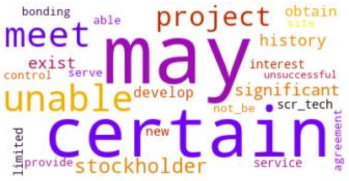
Topic #4 **Topic #5** **Topic #6**

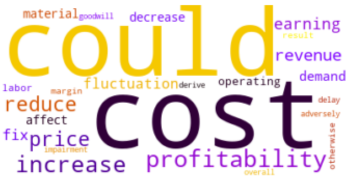


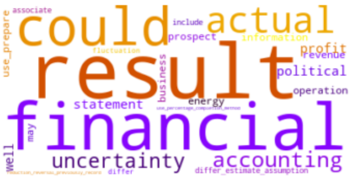




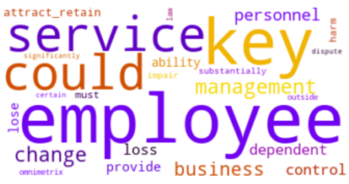
Topic #7 **Topic #8** **Topic #9**

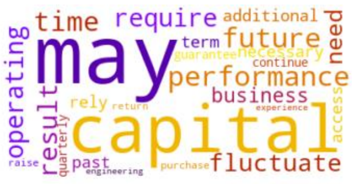


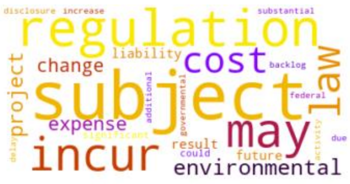




Topic #10 **Topic #11** **Topic #12**

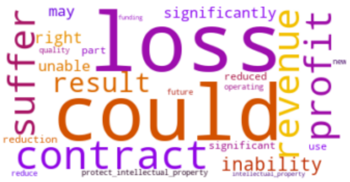




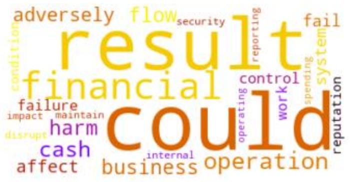


Topic #13 **Topic #14** **Topic #15**

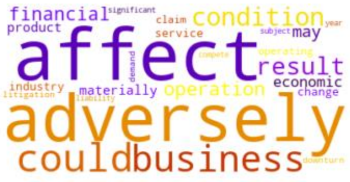
Table 10 Continued



Topic #16



Topic #17



Topic #18

While the generated topics do pick out some key words that can be attributed to different risk types, the key challenge here was identifying how to attribute (or map) each topic to one of the 18 unique risk types that were defined earlier. A lot of topics consisted of overlapping themes in terms of how a human user would interpret the risk types. The pyLDAvis library, which is used to visualization the results of LDA algorithm was further used to check the two-dimensional inter-topic distances between the obtained topics as seen in Figure 8. The visualization confirmed that a lot of topics were cramped in close proximity to one another, which explained the overlapping nature of the themes captured in them.

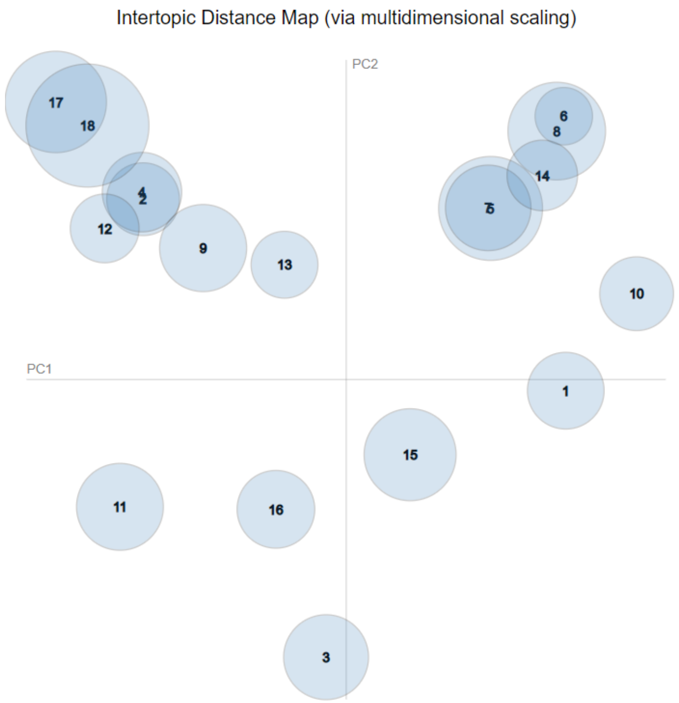


Figure 8 - Inter-topic distances plotted using pyLDAvis to show the two-dimensional distance between the 18 topics generated by LDA

The author experimented with some iterations of the LDA algorithm by tuning hyperparameters but the results were not very promising. The main challenge was that LDA allowed very little room for the author to incorporate the domain knowledge and the layout of the desired end result. At this stage, the need for a different (potentially more robust and state-of-the-art method) was felt which was eventually fulfilled by the implementation of the word-embedding algorithm FastText described in the next section.

4.3 Deep Learning and Text Classification

4.3.1 Gaps in traditional word count-based methods

As described in the previous section, the traditional word - count based methodology of implementing LDA algorithm was met with challenges. These traditional count - based techniques are simple and effective methods but due to the underlying nature of the methods, they lose information like the semantics, structure, sequence and context of the words in a textual format, which can be essential to classify text successfully. Additionally, count based models deal with words at an individual level which can lead to very large sparse word vectors and if the size of the data is not extremely large, the models developed tend to be poor, or may cause overfitting (blurring of boundary between noise and signal in data) due to curse of dimensionality (if data dimensionality is large, the volume of the space to store the data increases exponentially causing sparse data representations and hence causing difficulty in arriving at statistically significant analysis). During the stint of the author's graduate level coursework, he was introduced to some advanced and more powerful techniques of natural language understanding (NLU) and text classification which had the potential to overcome the challenges faced by the traditional methods.

4.3.2 Introduction to Word Embedding algorithms

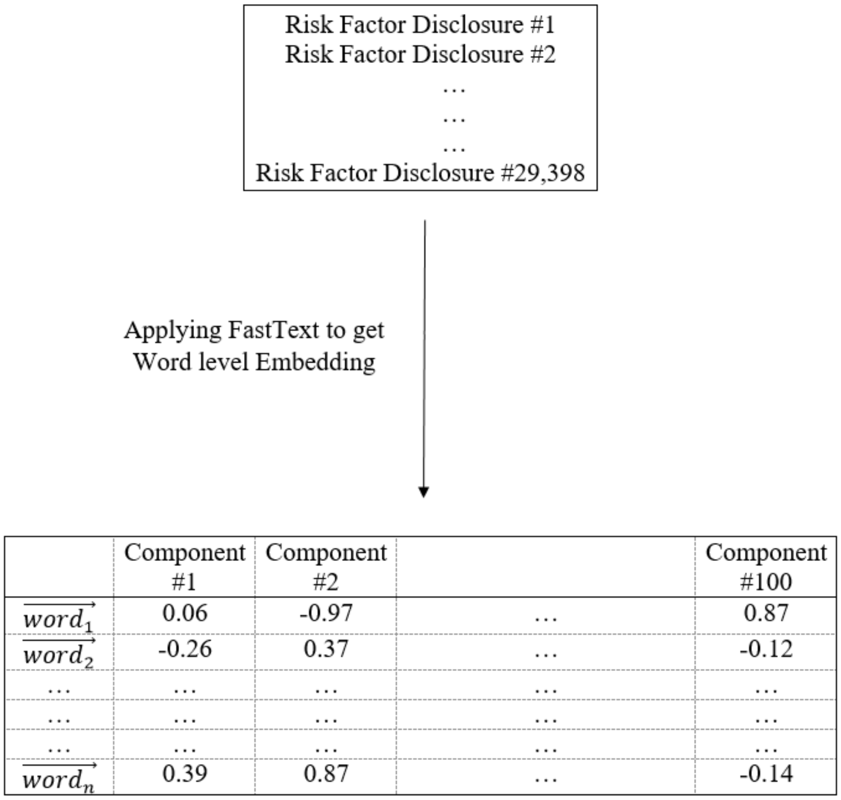
Over the last decade, a Deep Learning based set of language modelling and feature learning techniques called Word Embedding has gained popularity to overcome the limitations of the traditional text mining techniques. To understand the importance of some of the issues with the traditional methods listed above, consider the phrase, 'a word is characterized by the company it keeps'. Word embeddings are unsupervised models that can be applied to very large text corpus, to create a vocabulary of all possible words and be able to generate rich and dense word embeddings for every word in the vector space which represents that vocabulary. This framework facilitates understanding text by interpreting the words by their context, semantics and structure in a text. These methods usually allow the user to specify the size of the word embedding vectors as a hyperparameter. This facilitates the possibility of having a much lower dimension vector in comparison to the high-dimensional sparse vectors obtained using traditional count-based methods. The widely acclaimed word2vec model developed by Mikolov et. al. (2013) is able to generate embedding of unmatched quality at minimum cost, provided the size of the data is large enough. These word embeddings created by word2vec algorithm, have shown tremendous outperformance on many benchmark NLP tasks like syntactic and semantic word similarity, machine translation, or document classification.

4.3.3 FastText

Since the introduction of word2vec models in 2013, several improvements have been made by various computer scientists to come up with other techniques that perform better than word2vec in certain scenarios. For the present research, one such word embedding model known as FastText was implemented. The FastText model was developed by Facebook AI Research in 2016 as an extension to the vanilla word2vec model. The original FastText paper titled ‘Enriching Word Vectors with Subword Information’ by Bojanowski et al. (2016) provides details on how the methodology works. The word2vec model ignores the morphological structure associated with each word and only uses the word as a single entity. FastText model incorporates each word as a collection of characters, which is referred to as sub-word model in the paper. This framework goes deeper than just understanding words in the context, semantics and structure in a text. A sub-word is a collection of alphabets (or characters) that appear within the word. The paper recommends extracting all the sub-words of length for $n \geq 3$ and $n \leq 6$ where n is the number of characters in the sub-word. As an example, the word ‘apple’ can be considered a set of sub-words like {app, ppl, ple, appl, pple, apple}. In general, the FastText model allows a way to perform word embedding which is able to capture detailed and rich information in extension to word2vec models, and provides for a fast, robust and easy implementation to text classification and clustering applications. An added advantage that FastText allows for is that it can be used to obtain vectors for out-of-vocabulary words, by summing up vectors for its component sub-words, if at least one of the sub-words was present in the training data.

4.3.4 Implementation of FastText

The overall goal to use the FastText algorithm is to be able to attribute (or map) risk factor disclosures to one or more risk types identified in Table 8. To reiterate, those 18 risk types were obtained after a careful examination and content analysis of risk factor disclosures, and the overarching aim is to automate the process of applying a model-based classification approach to identify the risk factor disclosures with these risk types. To implement the FastText algorithm, a very popular open source natural language processing library in Python known as Gensim was leveraged. Gensim provides very good wrappers to use the FastText model available under gensim.models.fasttext module. All of the 995 risk factor files are read into Python and each file consist of multiple risk factor disclosures as discussed earlier. The unit of analysis is per risk factor disclosure. The first step is to apply the FastText algorithm to the entire set of 29,398 risk factor disclosures to obtain word embedding for each and every unique word (total n words) that appears in the corpus as illustrated in Figure 9. Each word vector is of 100-dimensions (default value in the Gensim library) where each dimension is a number that embeds information about the word, and its context and sub-words, which is generated using deep learning based techniques involving a large number of neural networks. At this stage, every possible word in the analysis has a word vector associated with it. (Note: Since this research was an unsupervised classification task of identifying important risk types from unlabeled textual risk disclosures, hyperparameter tuning was not of prime concern, as there did not exist "ground truth" labels to tune the model with. Hence, the author used the default parameters which has been developed by the creators of the FastText algorithm, and the successful validation exercise for this research presented in section 5.2 conducted with the help of subject matter experts, validates the model parameters too in-conjunction.)


Risk Factor Disclosure #1
Risk Factor Disclosure #2
…
…
…
Risk Factor Disclosure #29,398
Applying FastText to get Word level Embedding

	Component #1	Component #2		Component #100
$\overrightarrow{word_1}$	0.06	-0.97	…	0.87
$\overrightarrow{word_2}$	-0.26	0.37	…	-0.12
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
$\overrightarrow{word_n}$	0.39	0.87	…	-0.14

Figure 9 - Building the FastText Word Embedding Model

Once the word embedding for all the different words in the corpus are generated, the next step is to calculate the sentence vectors for the risk factor disclosure, which is referred here as the *source X*. To generate the sentence vectors for the *source X*, the mean sentence vector is obtained by averaging the vector for each word appearing in the risk factor disclosure (RFD) as shown in Figure 10. For example, selecting a risk factor disclosure from Table 7 presented earlier, one of the full risk factor disclosure sentences was *'We work in a highly competitive marketplace.'*, After pre-processing and text cleaning, the risk factor disclosure was reduced to *'work highly competitive marketplace'*. In this case, the sentence vector for this full risk factor disclosure would be an average of the word vectors of the constituent words '*work'*, *'highly'*, *'competitive'*, and *'marketplace'*. In this manner, the sentence vectors for each of the risk factor disclosures are obtained.

$$\overrightarrow{RFD_1} = \frac{\overrightarrow{word_1}+\overrightarrow{word_2}+\cdots+\overrightarrow{word_k}}{\mathrm{k}}$$

$$\overrightarrow{RFD_2} = \frac{\overrightarrow{word_1}+\overrightarrow{word_2}+\cdots+\overrightarrow{word_k}}{\mathrm{k}}$$

$$\cdots$$

$$\cdots$$

$$\cdots$$

$$\overrightarrow{RFD_{29,398}} = \frac{\overrightarrow{word_1}+\overrightarrow{word_2}+\cdots+\overrightarrow{word_k}}{\mathrm{k}}$$

where 'k' is the total number of words in each RFD

Figure 10 - Building the Vector Space for the Risk Factor Disclosures (RFD) (average of all word vectors in the sentence)

After all the sentence vectors for the *source X* are obtained, the next step is to generate the sentence vectors for the *target Y*, which represents all the 18 risk types (RT) that were identified by the content analysis procedure as listed in Table 8. It was seen that each risk type was determined with the help of certain keywords (or trigger words), that influenced the decision making during the content analysis process. Hence, to embed this valuable information in the model, the *target Y* vectors for each risk type are calculated by averaging the vectors for its associated keywords as illustrated in Figure 11. For example, the target vector for risk type ‘Accounting and Financial Reporting Risks’ is obtained by finding the average of the word vectors of its associated keywords *‘accounting’*, *‘accounting policy’*, *‘financial reporting’*, *‘financial statement’*, and *‘percentage completion method’* etc. (Note: Word vectors for phrases like *‘financial statement’* are obtained as average of word vectors of the words *‘financial’* and *‘statement’).*

$$\overrightarrow{RT_1} = \frac{\overrightarrow{keyword_1}+\overrightarrow{keyword_2}+\cdots+\overrightarrow{keyword_k}}{\mathrm{k}}$$

$$\overrightarrow{RT_2} = \frac{\overrightarrow{keyword_1}+\overrightarrow{keyword_2}+\cdots+\overrightarrow{keyword_k}}{\mathrm{k}}$$

$$\cdots$$

$$\cdots$$

$$\cdots$$

$$\overrightarrow{RT_{18}} = \frac{\overrightarrow{keyword_1}+\overrightarrow{keyword_2}+\cdots+\overrightarrow{keyword_k}}{\mathrm{k}}$$

where ‘k’ is the total number of keyword for each risk type

Figure 11 - Building the Vector Space for the 18 Risk Types (RT) (average of all keyword vectors for each risk type)

4.3.5 Text Classification using Cosine Similarity

The final step for the text classification is to determine a procedure to map the *source X* vectors to the *target Y* vectors. In text classification and similarity analysis, cosine similarity (Singhal 2001) is a very popular measure that is used to find the degree of similarity between two vectors. Given two non-zero vectors, $\vec{\boldsymbol{A}}$ and $\vec{\boldsymbol{B}}$, and the cosine similarity is derived using the Euclidean dot product formula as shown in Equation 1.

$$\cos\boldsymbol{\theta} = \frac{\boldsymbol{A}.\boldsymbol{B}}{||\boldsymbol{A}||\,||\boldsymbol{B}||} \tag{1}$$

The similarity values range from −1 (exactly opposite) to 1 (exactly same). It is a commonly used measure to evaluate the similarity between vectors arising from word embedding. The mapping of risk factor disclosure vectors to the target risk type vector(s) was done by using the Cosine Similarity measure between vectors as illustrated in Figure 12. The main idea here is that a target space comprising of 18 different vectors is created and each of the source risk factor disclosures are mapped to the target vector(s) that they most closely resemble.

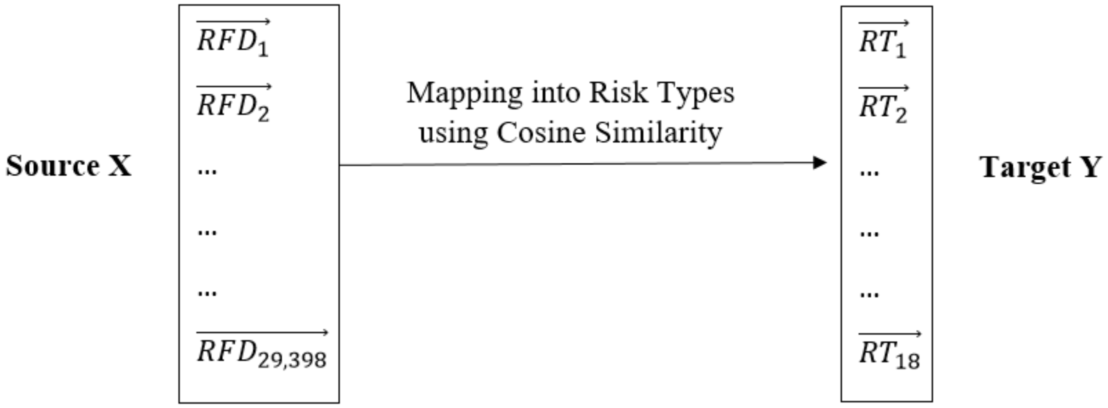


Figure 12 - Mapping the 29,398 Risk Factor Disclosure (RFD) vectors to the 18 Risk Type (RT) vectors using Cosine Similarity

4.3.5.1 Comment on Computation Time

Total time taken to calculate the word embedding on a desktop computer with 16 GB of RAM and an Intel Core i7 3.41 GHz processor, was 112 seconds. This step is the most computationally intensive task in this research, which in itself is achieved in less than 2 minutes. Hence computation time is not a significant factor in the process.

CHAPTER 5. DISCUSSION AND VALIDATION OF RESULTS

In the previous chapter, the details of implementation of text mining techniques to classify the risk factor disclosures was listed. All the extracted risk factor disclosures were mapped to the risk types by following a systematic procedure of applying word embedding algorithm FastText, followed by the cosine similarity measure. The research results and validation are presented in the following sections.

5.1 Research Results and Discussion

5.1.1 Distribution of the number of Risk Types identified for all risk factor disclosures

Figure 13 shows the distribution of the number of risk types identified for all the 29,398 risk factor disclosures. 14,196 risk factor disclosures (~48.3%) were classified as pertaining to only one risk type whereas 9,527 were classified as two risk types (~32.4%). Some risk factor disclosures were longer sentences which discussed several inter-related risks and were tagged with more than two risk types. At maximum, the highest number of risk types was found to be 6, with only 3 risk factor disclosures in the entire database, that fit this description. 406 risk factor disclosures (~1.3%) were not recognized as any of the risk type categories that the author wished to obtain. On examination of these 406 risk factor disclosures, it was found that most of them were either noisy data or just too vague to attribute to a specific risk. For example, risk factor disclosures, such as ‘there are many factors beyond the control of the company’ and ‘special note about forward-looking statements’, would fall in this category.

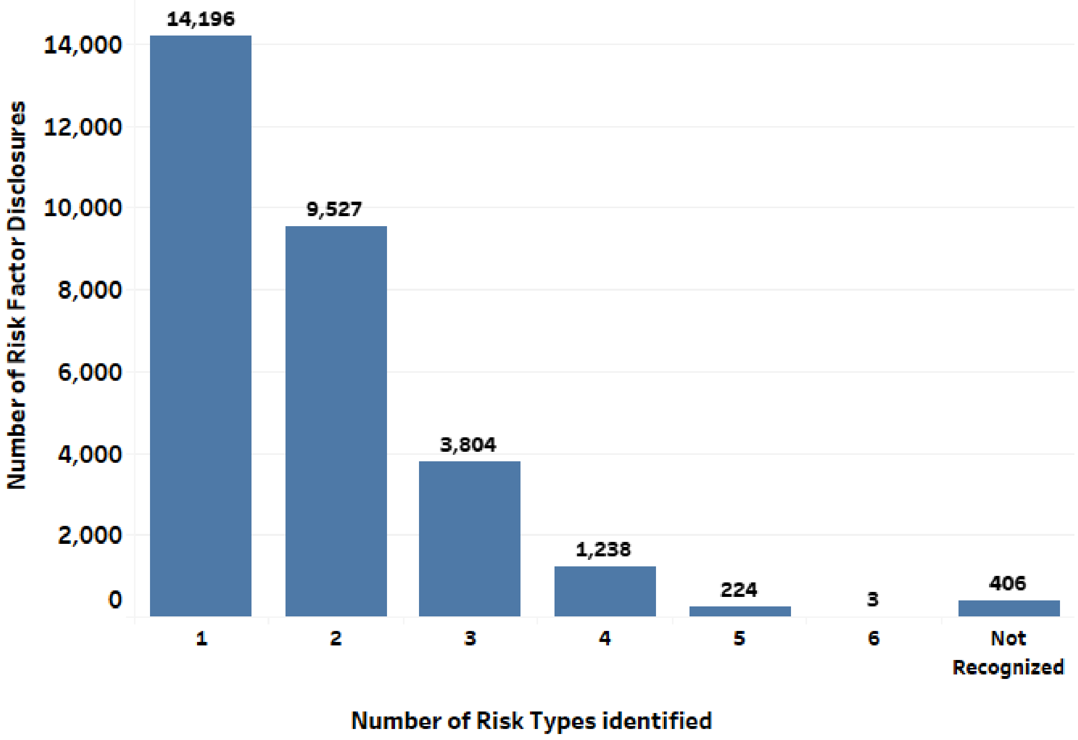


Figure 13 - Histogram of number of risk types identified for all risk factor disclosures

The assertion that longer sentences were identified with more risk types is supported by Figure 14 which shows the box-plot distribution of the number of risk types identified by the model versus the number of words in the risk factor disclosure. The box-plot was generated with the help of Seaborn and Matplotlib libraries in Python ("seaborn: statistical data visualization — seaborn 0.10.0 documentation," n.d.; "Matplotlib: Python plotting — Matplotlib 3.1.3 documentation," n.d.). Of the 14,196 risk factor disclosures that were identified with a single risk type, the median length of the sentence was found to be 17 words. A clear increasing pattern is observed in the plot, which shows that as the length of the risk factor disclosure sentences increases, they are identified with more risk types. Sentences which were identified with two risk types had a median of 24 words, the ones which were identified with three risk types had a median of 29 words. Similarly, the sentences which were identified with four, five and six risk types had a median of 32, 38 and 39 words respectively. In line with this finding, the 406 risk disclosures sentence which were not recognized by the model as any risk type, had the lowest median number of words at 11.

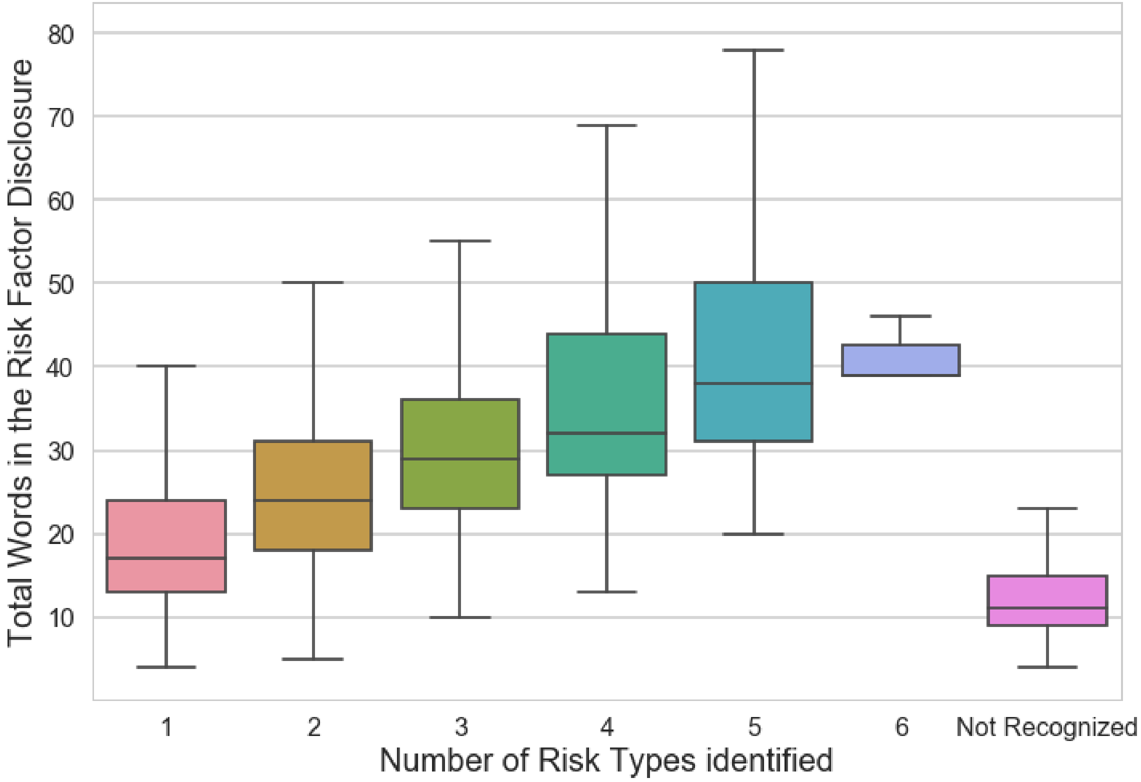


Figure 14 - Length of the Risk Factor Disclosure (no. of words in the sentence) and the number of Risk Types identified

5.1.2 Identification of the risk types at a firm - year level

It is impossible to present the results of all the 29,398 risk factor disclosures and each of their respective classified risk types individually, so a subset of results is presented for discussion. As a sample of the model-based classification, the 10-K file of Granite Construction for the year 2018 is utilized. In this filing, the company listed thirty-one (31) risk factor disclosures in the Item 1A - Risk Factors section. Table 11 depicts the risk types identified for all of these thirty-one risk factor disclosures. The developed model is able to check for attribution of all the 18 risk types independent from one another. It is similar to a binary classification (True/False or 1/0) over 18 target labels. It is seen that the various risk types associated for each of the risk factor disclosure of Granite Construction are identified. For example, the sentence "*Failure to maintain safe work sites could result in significant losses*" is correctly classified as 'Safety Risk'. Some other risk disclosures are associated to more than one risk type. For instance, "*Force majeure events, including natural disasters and terrorist actions, could negatively impact our business, which may affect our financial condition, result of operations or cash flows.*" is a risk factor attributed to both 'Natural and Manmade Disasters Risks' and 'Business Risks - Operational and Financial'. While going through the identified risk types for each risk factor disclosure, the reader is encouraged to refer to Table 8 which earlier presented a brief description of each risk type.

Table 11 - Risk Types identified for all Risk Factor disclosures of Granite Construction listed in the 10-K file for the year 2018

#	Risk Factor Disclosure *Identified Risk Type(s)*
1.	Unfavorable economic conditions may have an adverse impact on our business. *Macro Risks*
2.	We work in a highly competitive marketplace. *Competition Risks*
3.	Government contracts generally have strict regulatory requirements. *Governmental, Contractual and Regulatory Risks*
4.	Government contractors are subject to suspension or debarment from government contracting. *Governmental, Contractual and Regulatory Risks*

Table 11 Continued

5.	Our success depends on attracting and retaining qualified personnel, joint venture partners and subcontractors in a competitive environment. *Third Party Risks* *Human Resource Risks* *Competition Risks*
6.	Failure to maintain safe work sites could result in significant losses. *Safety Risks*
7.	As a part of our growth strategy we have made and may make future acquisitions, and acquisitions involve many risks. *Business Risks - Operational and Financial*
8.	An inability to obtain bonding could have a negative impact on our operations and results. *Business Risks - Operational and Financial*
9.	We may be unable to identify and contract with qualified Disadvantaged Business Enterprise (DBE) contractors to perform as subcontractors. *Third Party Risks*
10.	Fixed price and fixed unit price contracts subject us to the risk of increased project cost. *Cost Risks*
11.	Design-build contracts subject us to the risk of design errors and omissions. *Work Quality and Error Risks*
12.	Many of our contracts have penalties for late completion. *Time, Delay and Uncertainty Risks*
13.	Strikes or work stoppages could have a negative impact on our operations and results. *Time, Delay and Uncertainty Risks*
14.	Failure of our subcontractors to perform as anticipated could have a negative impact on our results. *Third Party Risks*
15.	Our joint venture contracts subject us to risks and uncertainties, some of which are outside of our control. *Third Party Risks*

Table 11 Continued

16.	Our failure to adequately recover on affirmative claims brought by us against project owners or other project participants (e.g., back charges against subcontractors) for additional contract costs could have a negative impact on our liquidity and future operations. *Third Party Risks* *Legal, Claims, Liabilities and Dispute Risks* *Business Risks - Operational and Financial* *Cost Risks*
17.	Failure to remain in compliance with covenants under our debt and credit agreements, service our indebtedness, or fund our other liquidity needs could adversely impact our business. *Business Risks - Operational and Financial*
18.	Unavailability of insurance coverage could have a negative effect on our operations and results. *Legal, Claims, Liabilities and Dispute Risks* *Business Risks - Operational and Financial*
19.	Accounting for our revenues and costs involves significant estimates. *Accounting and Financial Reporting Risks* *Cost Risks*
20.	We use certain commodity products that are subject to significant price fluctuations. *Cost Risks* *Macro Risks* *Financial Market Risks*
21.	We are subject to environmental and other regulation. *Governmental, Contractual and Regulatory Risks*
22.	Weather can significantly affect our revenues and profitability. *Natural and Manmade Disasters Risks*
23.	Increasing restrictions on securing aggregate reserves could negatively affect our future operations and results. *Business Risks - Operational and Financial*
24.	We may be required to contribute cash to meet our unfunded pension obligations in certain multi-employer plans. *Business Risks - Operational and Financial*

Table 11 Continued

25.	Force majeure events, including natural disasters and terrorists’ actions, could negatively impact our business, which may affect our financial condition, results of operations or cash flows. *Natural and Manmade Disasters Risks* *Business Risks - Operational and Financial*
26.	Changes to our outsourced software or infrastructure vendors as well as any sudden loss, breach of security, disruption or unexpected data or vendor loss associated with our information technology systems could have a material adverse effect on our business. *Technology Risks* *Third Party Risks* *Business Risks - Operational and Financial*
27.	Cybersecurity attacks on or breaches of our information technology environment could result in business interruptions, remediation costs and/or legal claims. *Technology Risks* *Legal, Claims, Liabilities and Dispute Risks* *Cost Risks*
28.	A change in tax laws or regulations of any federal, state or international jurisdiction in which we operate could increase our tax burden and otherwise adversely affect our financial position, results of operations, cash flows and liquidity. *Tax Risks* *Governmental, Contractual and Regulatory Risks* *Legal, Claims, Liabilities and Dispute Risks* *Business Risks - Operational and Financial*
29.	Our contract backlog is subject to unexpected adjustments and cancellations and could be an uncertain indicator of our future earnings. *Time, Delay and Uncertainty Risks* *Business Risks - Operational and Financial*
30.	Our business strategy includes growing our international operations, which are subject to a number of special risks. *Business Risks - Operational and Financial*
31.	Rising inflation and/or interest rates could have an adverse effect on our business, financial condition and results of operations. *Macro Risks* *Business Risks - Operational and Financial*

5.1.3 Risk types identified at a firm - year level

Once the individual risk factor disclosures are classified into the risk types, the next step is to develop a methodology to summarize them at a firm level. To obtain the risk profile for the firm-year observation, consider the example presented in Table 12. A hypothetical firm lists three risk factor disclosures (RFD) in its 10-K file for a specific year, and a total of four risk types are tagged, with risk type A being tagged twice, while risk type B and risk type C each being tagged once. The risk distribution for this firm - year observation is determined by a straightforward intuitive summarization to be 50% risk type A, 25% risk type B and 25% risk type C using Equation 2.

Table 12 - Example to illustrate the summarization of risk types at a firm-year level to obtain risk profile

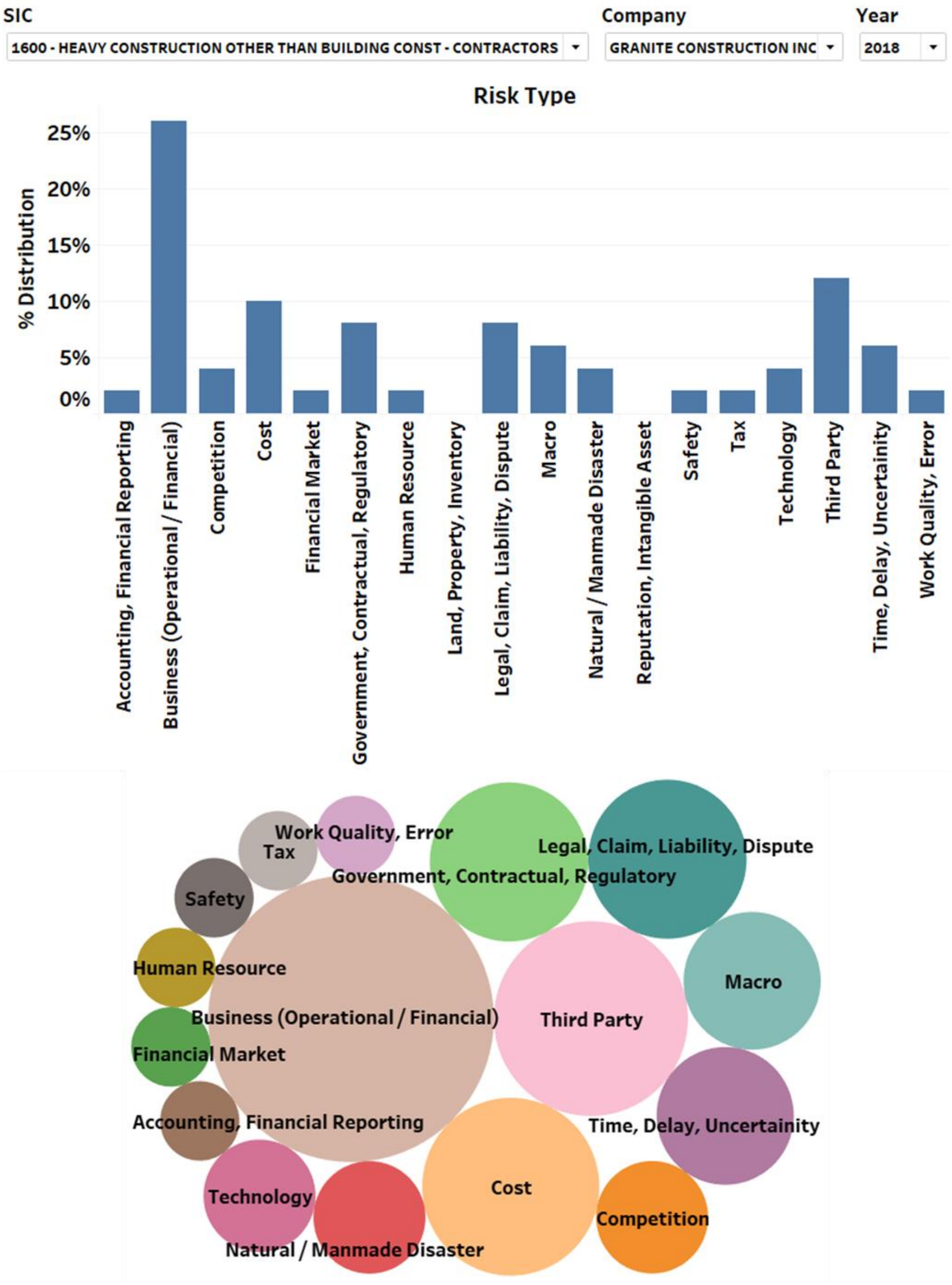
Risk Factor Disclosures	Number of Risks Identified	Label1	Label2
RFD_1	2	A	B
RFD_2	1	C	
RFD_3	1	A	

$$\% \; of \; Risk \; Type \; "X" = \frac{\sum Number \; of \; times \; Risk \; Type \; "X" \; is \; identified}{\sum Total \; number \; of \; Risks \; identified} * 100\% \qquad (2)$$

Equation 2 is applied to the risk profiles for all the firms for each of their available 10-K filing year. Figure 15 represents the distribution of risk types identified for Granite Construction based on their risk factor disclosures for the year 2018 in both histogram and bubble chart formats. It is noticed that about 25% of all the risk types are classified as

‘Business Risks - Operational and Financial’ which is anticipated to be found in the risk reports highlighting the financial health of a major construction firm. For example, the success of acquisition is a major business risk type for Granite Construction. In 2018, Granite Construction expanded its business operations by acquiring water management firm Layne Christensen Co., valued at about $565 million (“Finishing Acquisition of Layne Christensen, Granite Construction Revamps Roles | 2018-06-20 | Engineering News-Record” n.d.). The risk disclosure “*As a part of our growth strategy we have made and may make future acquisitions, and acquisitions involve many risks*.” represents this business risk, which has been correctly labeled as the business risk type by the algorithm.

There are several other stories related to Granite Construction in 2018 indicating to different types of risks affecting the company. For example, an ENR news article with the headline ‘Two Teams Certify Costs for Boston Green Line Extension | 2017-11-01 | ENR,’ n.d. mentions “The joint venture of Walsh Group, Bzarletta Construction and Granite Construction appears to have been unable to certify a cost estimate at or below the $1.3-billion cost limit to build the long-awaited Massachusetts Bay Transportation Authority’s Green Line Extension project in metropolitan Boston.” It serves as an example of how joint ventures and working in collaborative teams with third parties can lead to cost estimation challenges. The algorithm correctly categorizes the risk disclosures as ‘Third Party Risks’ and ‘Cost Risks’ which emerge as the second and third most commonly tagged risk types for Granite Construction in 2018. Some other prominent risk types identified are ‘Legal, Claims, Liabilities and Dispute Risks’ and ‘Governmental, Contractual and Regulatory Risks’ which a large company like Granite Construction, that has a significant portion of its revenue from its public sector clients, is expected to be mindful of.



SIC: 1600 - HEAVY CONSTRUCTION OTHER THAN BUILDING CONST - CONTRACTORS

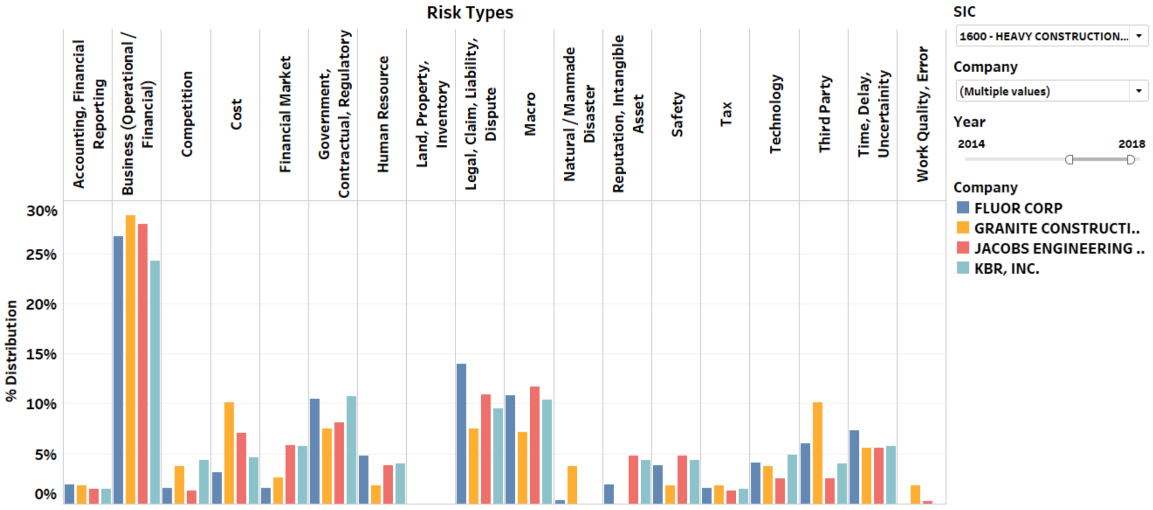
Company: GRANITE CONSTRUCTION INC

Year: 2018

Figure 15 - Distribution of risk types identified for Granite Construction based on their risk factor disclosures of the year 2018

5.1.4 Comparison of risk profile of four companies in SIC group 1600 over a five-year period

Comparison of risk profiles for companies is an important benchmark for different stakeholders in the construction industry to make business decisions. The Tableau dashboard developed allows to compare the distribution of risk type between various companies over a period of time. For the purpose of demonstration, Figure 16 shows the comparison of risk types identified for four companies, i.e. Fluor Corp, Granite Construction, Jacobs Engineering and KBR, within the SIC group 1600 - 'Heavy Construction other than Building Construction - Contractors' over the time period 2014-2018. As these firms belong to the same sub-group in the construction industry, a comparison between their risk profiles is intuitive and has merit. All companies typically perform some form of risk profile comparison with its prominent competitors.



SIC
1600 - HEAVY CONSTRUCTION...
Company
(Multiple values)
Year
2014
2018

Figure 16 - Comparison of risk types identified for Fluor Corp, Granite Construction, Jacobs Engineering and KBR based on their risk factor disclosures for the years 2014 - 2018

It is observed that the profiles of the distribution of the risk types discussed in the 10-K filings for these companies, are largely similar. 'Business Risks - Operational and Financial' are found to be most prominent for all the companies but Granite Construction's

10-K filings discuss it more often than KBR. Fluor Corp depicts ‘Legal, Claims, Liabilities and Dispute Risks’ and ‘Time, Delay and Uncertainty Risks’ more often than its competitors whereas Jacobs Engineering lists a higher frequency of risks which are identified as ‘Macro Risks’. Similar comparisons are possible for different groups of companies belonging to other sub-industry groups over the desired time-frame.

5.1.5 Comparing risk profiles among different sub - industry groups over a five-year period

There are eight SIC codes for different subsectors of the construction industry as summarized earlier in Table 4. A comparative assessment of risks was conducted for these different sub-industries. Figure 17 shows a stacked bar chart of the distribution of all the risk types identified for each of the SIC group over the time period 2014-2018.

Overall, most risk types have similar profiles across the different sub-sectors of the construction industry. However, some differences were observed. On closer examination of ‘Safety Risks’, it was found that compared to any other sub-sector, it had the highest percentage of occurrence for SIC group 1600 - ‘Heavy Construction other than Building Construction - Contractors’. This indicates that since Heavy Construction typically involves increase in complexity and uncertainty in the nature of the projects, and often uses non-standard equipment and engineering approaches, ‘Safety Risks’ become an even more important consideration.

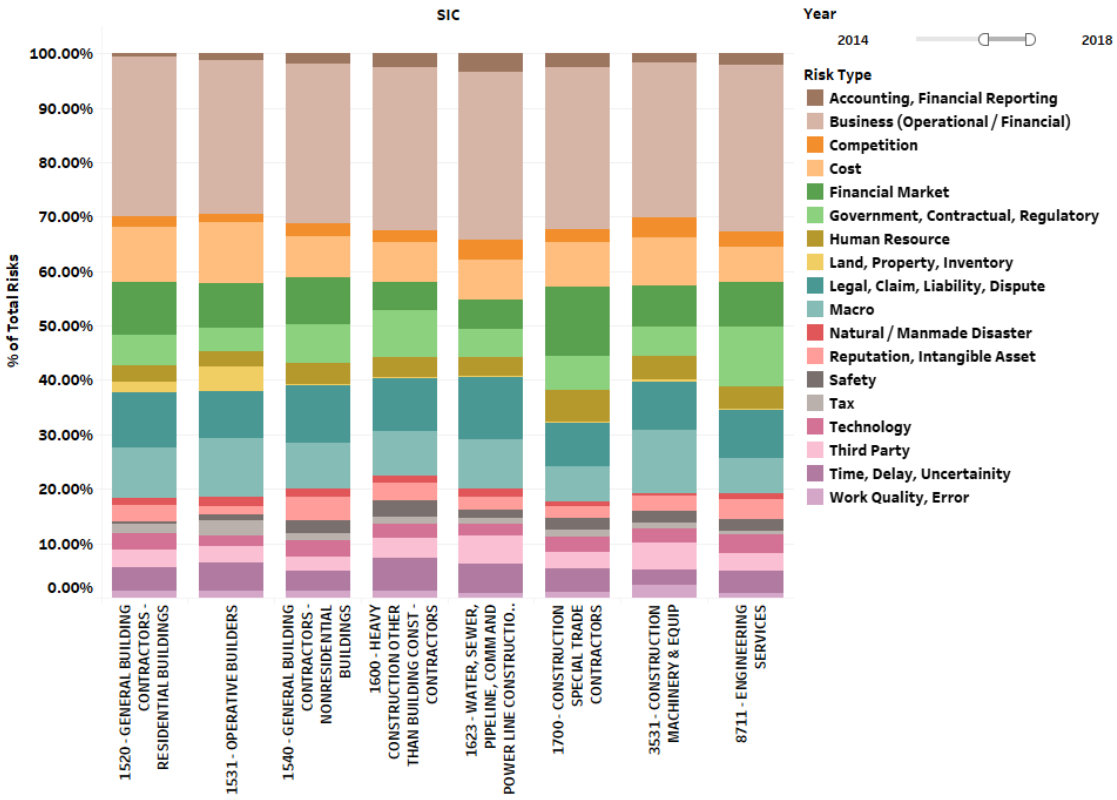


Figure 17 - Comparison of risk types identified for different SIC groups for the years 2014 - 2018

‘Time, Delay and Uncertainty Risks’ were least frequently tagged for SIC group 3531 - ‘Construction Machinery & Equipment’ compared to every other SIC group that comprises of companies involved in the design and construction itself. This can be the case as machinery and equipment companies tend to have more streamlined and automated manufacturing units, which increasingly use lean principles, and do not have to face the same levels of delays and change orders like contractors working on construction sites.

5.1.6 Increasing number of risk factor disclosures over the time period 2006-2018 for all the public companies in the construction industry

So far, the discussion on risk type identification and distribution was limited to a company or a sub-industry level. Figure 18 presents a look into the average number of risk factor disclosures in the 10-K files for all construction companies over the years. An increasing trend is observed since 2006 when the number was close to 20, while it has risen to just over 35 risk disclosures per file in 2018. This is a very interesting trend that can potentially point to a combination of factors, a non-exhaustive list of which includes: -

- Increase in risk faced by construction companies over the time-period
- Better identification and acknowledgement of risks in the forward looking 10-K files provided by the construction companies
- The companies improving the thoroughness and the richness of the risk factor disclosure in response to SEC enforcing strict regulatory requirements and threat of penalties for vague or non-informative disclosures
- Equity analysts in investment banks and financial institutions placing an increasingly significant scrutiny on the 10-K filings of public companies and thus a poor-quality filing leading to a negative equity rating assigned to the company, potentially causing a decline in its stock price
- Rise of systematic (or algorithmic) trading in the financial markets and their adoption of alternative sources of data, such as textual SEC filings, leading to even more emphasis on the 10-K files.

A couple of examples of specific risks that were found to have originated during the time period of analysis are listed here. It was observed that risks related to climate change and its impact on business and regulations were first discussed in the 10-K files of construction companies in the year 2010 and since then they have been discussed regularly. Another example is risks due to cybersecurity threats which first started popping up in 2012-13 and have seen a growth in their incidence since then. The overall increasing trend in risk factor disclosure reinforces the usage of the SEC 10-K filings data as a rich source of information for continued future research too.

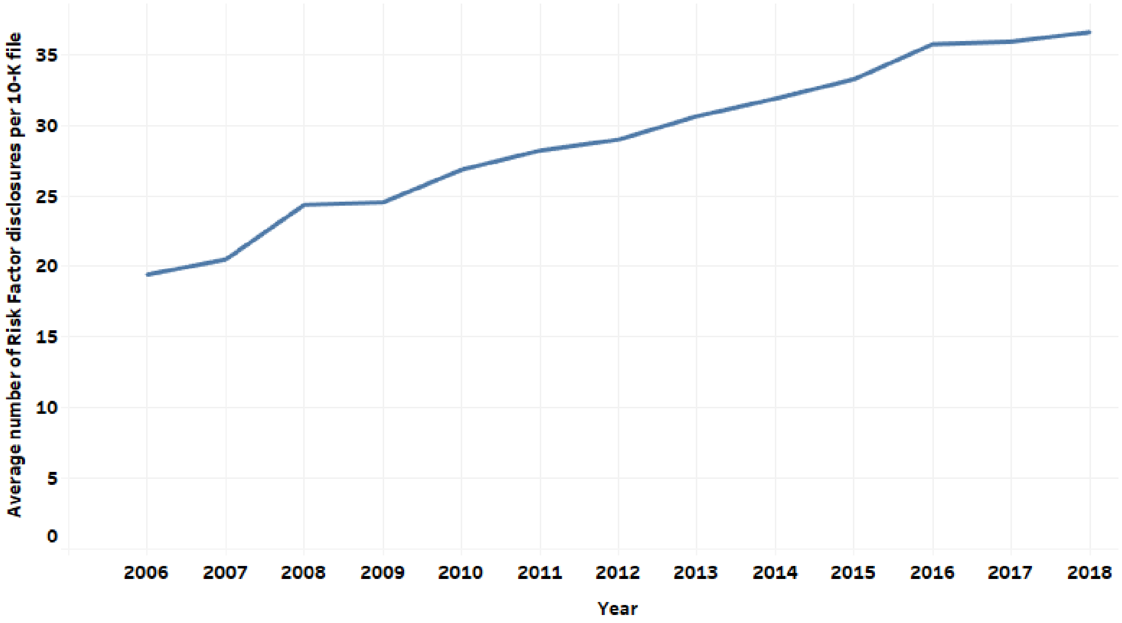


Figure 18 - Average number of risk factor disclosures per 10-K file over the years for construction industry

5.1.7 *Evolution of risk types over the time period 2006-2018 for all the public companies in the construction industry*

Figure 19 depicts an area chart of the overall trends observed in the risk type distribution for all the public companies in the construction industry over the time period of 2006-2018. Note that this chart is a massively aggregated representation with each year normalized to 100% to facilitate comparison among different years. The proportion of risk types are stable over time with slight increase or decrease in some of them found in the risk factor disclosures. All the risk types are important items for construction industry and are expected to be disclosed and addressed in the 10-K filings for the companies throughout the time-period.

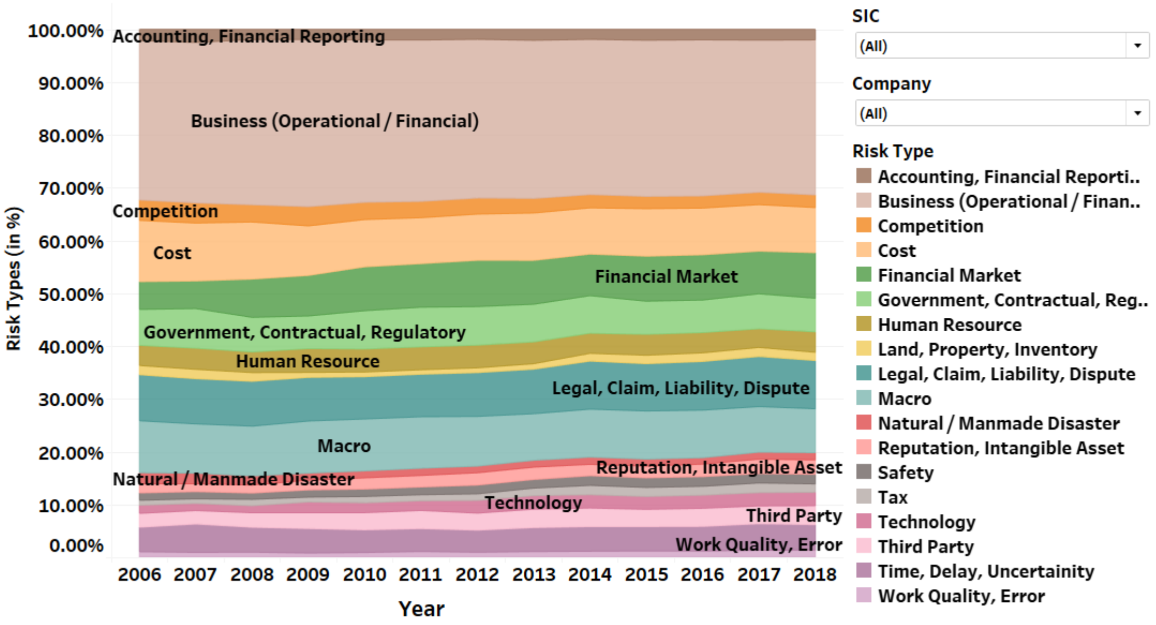


Figure 19 - Evolution of trends of all the 18 risk types over the time period 2006-2018 for all the public companies in the construction industry

5.1.8 Emerging trends of certain risk types over the time period 2006-2018 for all the public companies in the construction industry

Figure 20 highlights four risk types that are found to have a general upward trend for the overall time-period of 2006 - 2018. It was recognized that the ‘Technology Risks’ has found increasing mention in the 10-K files over time. This can be attributed to the increased adoption of BIM, Virtual Design and Construction (VDC) and Internet of Things (IoT) in the construction industry and the rational focus of assessing risk associated with newer technologies. With the increase in storage of project data on cloud and migration to online modes of communication, cybersecurity threats have become a significant factor in the modern construction ecosystem as well ("Construction Cybercrime Is on the Rise | 2019-05-08 | Engineering News-Record," n.d.).

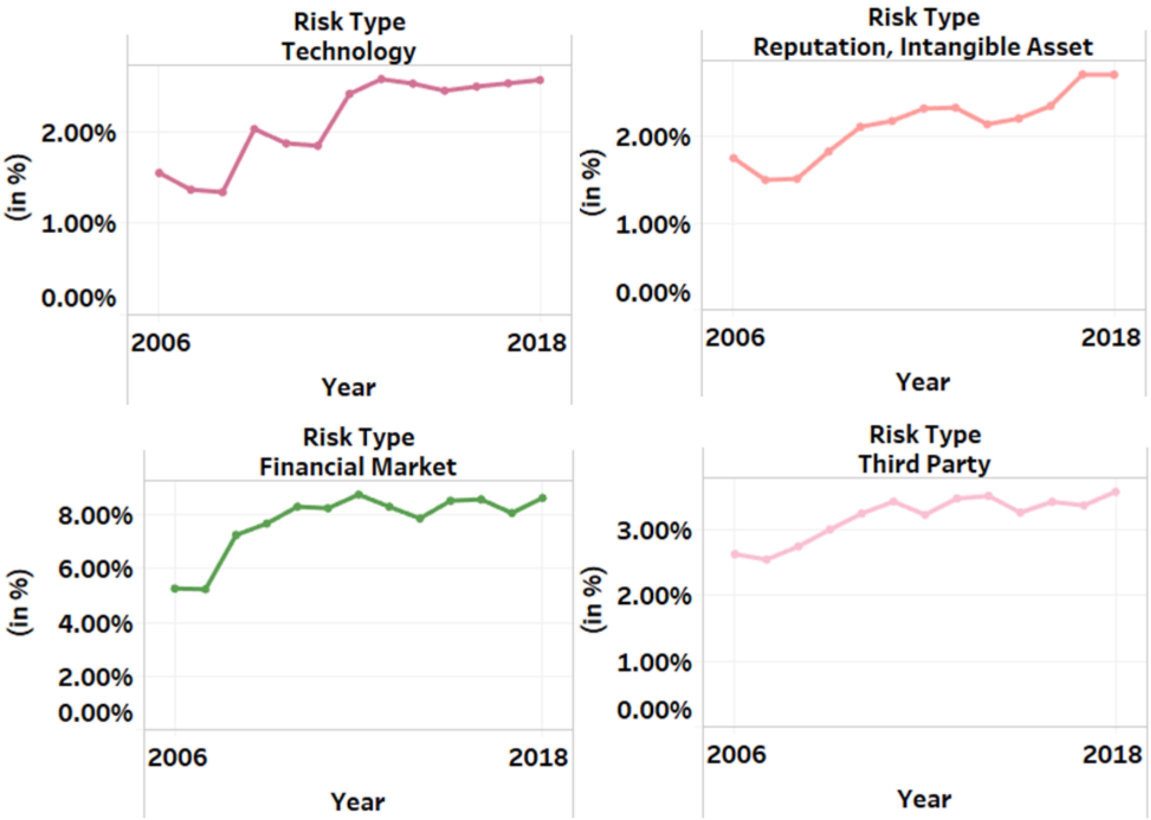


Figure 20 - Trend plots of certain risk types over the time-period 2006-2018 for all the public companies in the construction industry

'Reputational and Intangible Risks' have also found higher incidence in these risk disclosures, as these types of risks have become increasingly important in the past few years with the spread of social media and other factors, thus companies becoming more and more cognizant of how their brand is being perceived by the consumers and other stakeholders in the industry that it does business with. Leonard (2018) highlights that since social media has provided a voice to all, corporations are vulnerable to potential reputation damage, irrespective of it being based on fact or fiction. It is further stated that a lowtolerance for even minor wrongdoings and deep-rooted anti-corporate sentiments among some people can fuel a barrage of negative social media conversations potentially leading to full-blown crises. For instance, in 2017, United Airlines forcibly removed a passenger from overbooked plane which was captured on video and which led to severe outrage on social media websites like Twitter and Facebook, caused their stock to plummet and therefore the market capitalization of United Airlines to drop by $1.4 Billion ("United Airlines: Stock Drops Following Passenger Incident in Chicago | Fortune," n.d.). This serves as an example of how quickly things can go south if a firm is negligent with reputational risks.

The discussion of 'Financial Market Risks' witnessed a significant spike during the years 2008-09 which is in line with the great economic recession of the last decade, and has remained an important consideration for the companies in the construction industry. Project delivery methods such as Design - Build has moved from alternative method to becoming a mainstream approach as per a 2018 Design-Build Institute of America (DBIA) study which suggests that nearly half of all projects nationwide will be delivered by design-build approach ("Design-Build Moves from Alternative to Mainstream – DBIA," n.d.). Other delivery methods such as integrated project delivery (IPD), and innovative financing methods through the use of public-private partnership (P3) have also emerged which have necessitated increased collaboration among stakeholders and has led to transfer of risks from owners to other entities. This can be attributed to the increase in 'Third Party Risks' (e.g., risks associated with designers, sub-contractors, and other partners in joint-ventures).

5.2 Validation of Results

The presented research has two layers of findings, first of which is at the level of individual risk factor disclosures extracted per company per year and its associated risk type classification identified by the developed model, and secondly the summarization of the risk identification results at broader levels to observe various risk trends and patterns. The validation methodology adopted addresses both these cases as detailed in the following sub-sections.

5.2.1 Survey of Human Subjects

To ensure that the developed text mining model is able to appropriately read and synthesize the individual risk factor disclosures and correctly identify the associated risk types, a human subject validation test was employed. Text mining research in construction industry has numerous examples of validation of the obtained model and results with the help of the subject matter experts (SMEs). Carrillo et. al. (2011) applied a procedure of knowledge discovery from text mining on 48 post-project reviews of two construction firms and they compared the findings to the key themes identified by a manual analysis. Their findings were evaluated by presentation to the companies though project meetings. Tixier et. al. (2016) developed a content analysis tool for implementation in construction safety, to identify the causes and outcomes from injury reports. They surveyed seven researchers to provide feedback on their results by using expert opinion as a baseline to judge the precision and recall of their text mining model. Lee et. al. (2019) developed an automatic text mining model to extract poisonous clauses using rule-based NLP to scan construction project contracts. They evaluated the model with the help of seven domain experts in contracts and claim management.

In most text mining applications in the construction industry research (including the present one), one of the main contributions of the developed methodologies is that it is able to capture the domain knowledge of a human user and mimic human interpretation and judgement of natural languages, to accomplish a task which would be impossible for a human user to conduct manually. A general trend that has been observed by the author during his study of literature and his knowledge acquisition of text mining methods and NLP is that, the most important validation test (and sometimes the only test) for these types of techniques is whether the results make semantic and logical sense from a human user perspective. Another way to re-state the previous sentence would be, the key check for text mining methods is often to examine whether the developed model is able to perform the text classification on similar lines to what a human subject with appropriate domain knowledge would carry out.

To validate the results of the developed text classification model used to identify and classify the risk factor disclosures with one or more risk types, a human subject validation survey was conducted. Each survey respondent was given a small subset of randomly sampled risk factor disclosures to classify into. Since there are 18 different risk types identified and used in this research, after a couple of pilot surveys under the author's supervision, it was observed that the survey respondents had a tough time keeping up with each of the 18 risk types and their definitions, which had to be used in labelling the risk factor disclosure sentences. And since majority of the survey respondents had to be surveyed over email, the need to make the process easier for the respondents was felt. To solve this issue, the survey was simplified and restricted to include risk factor disclosures that were identified with either one or two risk types, and each survey respondent was given a total of six risk types and their definitions to choose from. The ultimate goal was to see if the risk types that these respondents choose after reading the risk factor disclosure would line up with the risk types identified by the automated text mining model.

It was important to establish the basic requirements for the survey respondents. In order to do that, author referred to the guidelines provided by the SEC on the ‘Item 1A - Risk Factors’. As mentioned earlier in section 3.2, the SEC mandates that the risk factors section must be written using plain English principles that include short sentences, definite, concrete and everyday words with no legal or highly technical jargon and no multiple negatives. The goal is to make sure that the contents are understood by a layperson as well. With this requirement in mind, for the present research, the list of people surveyed included graduate level students with relevant educational backgrounds, industry professionals working either in the construction industry or the finance industry, and also a professor at a research university. All of the people surveyed easily met the basic requirements laid down by the SEC. Table 13 presents a brief background of the survey respondents, along with the results from the survey. A total of 18 people were surveyed and their responses on the risk types identified were tallied with the model generated risk classifications, to observe the agreement rate with the model. The survey respondents were given the option to skip any risk factor disclosure that they may not be able to understand or weren’t sure about.

Table 13 - Brief backgrounds of the Survey Respondents and the Survey results

Occupation	**# of Risk Types identified**	**# of Risk Types that agree with the model**	**Agreement Rate (in %)**
Professor of Construction Engineering	20	18	90.00%
Senior Managing Director, at an Infrastructure Consulting firm	20	18	90.00%
Director Project Management, at a Healthcare Construction firm	38	30	78.95%
Quantitative Analyst, at major US Bank	38	32	84.21%
Structural Engineer at globally reputed Structural Design firm	32	31	96.88%
Finance Professional in a Construction firm	27	23	85.19%
Transportation Analyst in Public Sector body	20	18	90.00%
Professional at Environmental Consulting firm	32	24	75.00%
Forensic and Litigation Consultant, at a Legal Construction Consulting firm	33	26	78.79%
Assistant Vice President of Financial Forecasting, major US Bank	20	18	90.00%
Project Engineer at Construction Engineering firm	16	13	81.25%
Current student, MS in Economics at Paris School of Economics	29	25	86.21%
Current PhD student in Civil Engg at Georgia Tech	20	18	90.00%
Current PhD student in Building Construction at Georgia Tech	20	18	90.00%

Table 13 Continued

Current student, MS in Building Construction at Georgia Tech	20	16	80.00%
Current student, MS in Building Construction at Georgia Tech	19	17	89.47%
Current student, MS in Building Construction at Georgia Tech	18	17	94.44%
Current student, MS in Building Construction at Georgia Tech	20	19	95.00%

Overall, it was observed that the in all the surveys conducted, the agreement rate with the model generated labels was high. On aggregate, 442 risk labels were identified by all respondents, out of which 381 were the same as the model generated labels, which led to an overall agreement rate of 86.19%. Considering that each survey respondent had six labels to choose from, a random choice of labels would have led to an agreement rate of 16.67%. Therefore, an overall agreement rate of 86.19% shows that the respondents by and large agreed with the model generated labels. The author closely examined some of the cases when the survey respondent's label did not match with the model generated label and noticed that the reasons for the discrepancies were largely due to subjective nature of some of the risk factor disclosures which can be interpreted differently by different people, based on their respective experience and backgrounds. For example, the risk factor disclosure, *'A terrorist attack or the threat of a terrorist attack could have a material adverse effect on our business.'* was identified as a 'Natural and Manmade Disasters Risks' by the model, but the survey respondent chose to label this as a 'Macro Risk'. Another risk factor disclosure, ‘*If we are unable to protect our intellectual property adequately, the value of our patents and trademarks and our ability to operate our business could be harmed.*’ was identified as ‘Reputational and Intangible Risks’ by the model but was labelled as ‘Legal, Claims, Liabilities and Dispute Risks’ by the survey respondent. These few cases of disagreements of survey respondent’s classification with the model generated classification were not found to be a significant deterrent and the overarching results were validated by the survey takers. An important note here is that the model allows to be tailor-built for any user, or company’s interpretation, such that the outcome can be tweaked to include risk definitions and interpretation as determined by the use case. The present version of the model is built with the domain knowledge of the author with the help of inputs from his doctoral adviser, and thus reflects their judgement and interpretation.

5.2.2 Comparison with existing literature

Figure 21 shows the comparison of the ‘Macro Risks’ identified by the developed model for the construction industry with the Macroeconomic Risk identified by Bao and Datta (2014) for the time period of 2006-2010. As earlier mentioned in the literature review, Bao and Datta (2014) applied a sent-LDA topic modeling algorithm to risk factor disclosures obtained from 10-K files of all public companies (not just construction industry), between the time period 2006-2010. In that paper, the authors highlight the Macroeconomic risk count identified for the same time-period as one of their discussion points. The information obtained from a figure in that paper is approximately recreated here to compare the trend with the percentage of risk types identified as ‘Macro Risks’. It is acknowledged here that because of the difference in the scope of the underlying data (all companies versus only construction companies etc.), the comparison is not expected to be exactly same, but the similarity in the trend is definitely noteworthy. Bao and Datta (2014) explain how the incidence of Macroeconomic Risk significantly increases in the disclosures for the year 2009, which they attribute to financial crisis of which occurred in late 2008, as a possible factor. An analogous trend is observed in the current research for the construction companies too, which provides the closest comparison that was possible to make, with a text mining-based study applied to the SEC 10-K financial filings.

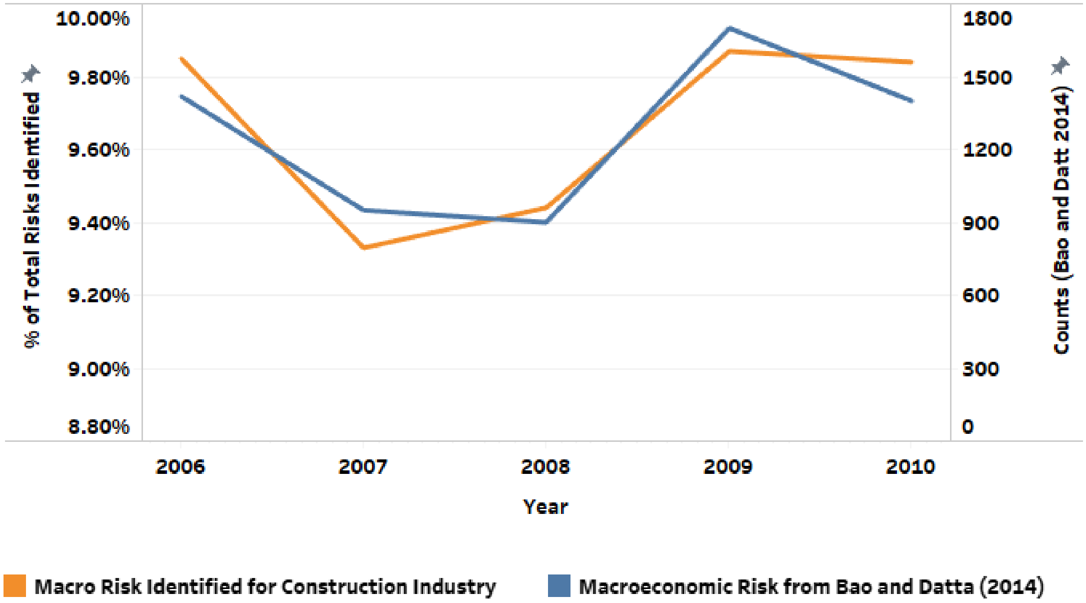


Figure 21 - Comparison of Macro Risks identified with Bao and Datta (2014) over the time-period 2006-2010

Finally, in addition to the similarity in the trend for Macro Risks, Bao and Datta (2014) also state that the 10-K files that they used had an average of 21.78 risk factor disclosures per file. Figure 18 touched upon the average number of risk factor disclosures by year for the present research. When the time period is filtered to 2006-2010 and all the 10-K files for this period considered together, the average number of risk factor disclosure is found to be 23.34 per file, which is comparable to the findings of Bao and Datta (2014). The numbers are of course not expected to be exactly equal because the underlying data set is not completely the same (all companies 10-K files versus only construction companies 10-K files) but the numbers are close to each other, which adds another layer of validity to the process of risk factor extraction from the 10-K files.

CHAPTER 6. CONCLUSIONS, LIMITATION AND FUTURE WORKS

Risk management in the construction industry has been a topic of prime interest since time immemorial. It is directly tied to the financial success of construction project but in addition to that, it is also tied to the success in terms of quality of the project delivered, the safety of the workers involved, protection from claims and liabilities, schedule delay and extension avoidance, compliance with the several laws that govern the industry etc. Practitioners in the industry and researchers in academia have given a very important share of their time and effort to study and make progress in identifying innovative practices and approaches to mitigate risks, and rightfully so, given the potential disastrous implications that are associated with instances of ill-management of risk. As a large industry, it is extremely important to keep up with the dynamic change of the impact and influence of various construction industry wide risk factors.

This research study presents a new and innovative approach of carefully examining and analyzing the risk disclosures made by public construction companies in their annual SEC filings. Literature review of risk studies in the construction industry indicate that although there have been quite a few studies on a task level, project level and program level; there exists a gap in the investigation of various risk types on a macro level, i.e. the enterprise level, sub-industry level and the industry level. Challenges to macro level studies are identified which include the extremely large scope of such a research coupled with difficulty in conducting traditional survey type studies.

A state-of-the-art methodology is implemented which applies advances in deep learning and natural language processing to systematically identify and classify risk types from unstructured textual risk factor disclosures made by public construction companies in their annual SEC 10-K filings. The current study introduces SEC 10-K filings as a new source of data in the domain of construction research which is a dynamically enriching, professionally audited, reputed and widely examined source of information. The word embedding algorithm FastText and Cosine Similarity measure are leveraged to create a risk classification model that is able to successfully attribute the textual risk factor disclosures to 18 different risk types. Results are presented in conjunction with support from literature and construction industry news articles. Influence of significant external events, such as the financial crisis of 2008, is observed in the trend patterns of the risk factor disclosures and the study highlights some of the risk types which are observed to have an increasing trend over the time-period 2006 - 2018. The methodology of the research is validated with the help of a structured survey of industry professionals along with evidence from literature.

The study bridges the difficult challenge of keeping a track of risk types affecting various construction companies and sub-sectors within the industry. The quality of the risk factor disclosure for the construction companies can improve if there is a continued close scrutiny on them and the current research pushes the envelope in that regard as well. The developed model serves as risk thermometer to identify important patterns and trends, both over cross-sections covering a multitude of companies, and over a time-series of several years. This research helps in reducing any existing information asymmetry by introducing a new and unique research methodology that combines knowledge from construction management, finance and artificial intelligence. The methodology developed is readily transferrable to other text classification use-cases in the construction industry.

The findings of the study have practical implications for professionals and researchers who study risk in the construction industry. It is anticipated that public construction companies which file 10-K reports can gain good insight from the study by understanding the behavior of their peers and the industry as a whole. Private construction companies who are not mandated to file risk disclosures with the SEC, can also benefit by scrutinizing the risk profiles of similar public companies that submit 10-K filings. Risk assessment for partnering decisions between firms to form teams for joint venture projects can also be augmented. Banks and capital financing institutions can use the findings to supplement their risk management models. Insurance companies can use the identified risk information for developing better risk pricing models. Investors in the construction industry can make more informed decisions about the major types of risk affecting the financial bottom line of their business. The pricing of risk into construction contracts can be made more efficient and cost-effective as the research can add previously unknown insights to the process. Overall, it is expected that the research will add meaningful contributions to all the stakeholders involved in the construction industry and research, by providing an additional way of examination of industry wide risk factors.

Every research study has its own limitations and the present study has certain limitations as well. The identification of sub-industry and industry level trends is limited to publicly traded construction companies only because no standard data source exists which captures similar information of privately-owned construction companies. Another important limitation is that the risk factors disclosed in 10-K filings, although professionally audited, are done by the companies themselves, which can lead to possible biases. Firstly, some companies may choose to withhold information if disclosure of certain risks can be damaging to their business and reputation. On the other hand, because of potential penalties and liabilities for the failure to disclose important risks, some companies may elect to disclose unnecessary and unimportant risk factors as well, just to protect themselves from any possibility of problems and issues in this regard. Finally, the disclosed risk factors in 10-K files do not cover any risk that the company itself is not cognizant of, but in reality, is a major factor. This can provide a false sense of security around the results, as it does not directly reflect risks unknown to the company. Hence, these issues need to be acknowledged before using the results of the present research. Research undertakings such as the present work and any future studies on the 10-K filings serve as a step towards increasing scrutiny on them, and can be expected to improve the overall quality of the filings.

As far as future works are concerned, this study can lead to several types of future research works depending on the practical needs that can arise. The results of the risk identification and assessment model can serve as a building block to study the inter-dependence of risk types with one another (for example, if the occurrence of one type of risk is found to be increasing, how does it affect some other risk type etc.). With continued advancements in the field of NLP and Machine Learning, newer methods can be explored that can provide more ease and flexibility in determining important risk types. The present work with appropriate financial support, human resource and survey of the existing needs in the market (market research), can be converted into a commercial risk identification and assessment product. Overall, the methodology is applicable to several text classification problems in the construction industry and inspired works are encouraged.

APPENDIX A. LIST OF ALL FIRMS BY SIC GROUPS

SIC 1520 - General Building Contractors - Residential Buildings

All American Group Inc

Brookfield Homes Corp

Extensions, Inc.

Fortune Brands Home & Security, Inc.

Global Diversified Industries Inc

Installed Building Products, Inc.

Lennar Corp /New/

Pernix Group, Inc.

Prospect Global Resources Inc.

Select Interior Concepts, Inc.

TOUSA Inc

UNR Holdings Inc

VRDT Corp

SIC 1531 - Operative Builders

Ashton Woods USA L.L.C.

Av Homes, Inc.

Beazer Homes USA Inc

Calatlantic Group, Inc.

California Coastal Communities Inc

Century Communities, Inc.

Comstock Holding Companies, Inc.

Dominion Homes Inc

Green Brick Partners, Inc.

Heavenstone Corp

Helpful Alliance Co

Horton D R Inc /De/

Hovnanian Enterprises Inc

Kb Home

Kimball Hill, Inc.

LGI Homes, Inc.

M I Homes Inc

MDC Holdings Inc

Meritage Homes Corp

New Home Co Inc.

NVR Inc

Orleans Homebuilders Inc

Ryland Group Inc

Shea Homes Limited Partnership

Stanley-Martin Communities, Llc

Taylor Morrison Home Corp

Toll Brothers Inc

Tri Pointe Group, Inc.

UCP, Inc.

WCI Communities Inc

William Lyon Homes

SIC 1540 - General Building Contractors - Non-residential Buildings

Alternate Energy Holdings, Inc.

Aquentium Inc

Continental Cement Company, L.L.C.

Liandi Clean Technology Inc.

Servidyne, Inc.

Sports Field Holdings, Inc.

Summit Materials, Llc

Tutor Perini Corp

Vicapsys Life Sciences, Inc.

SIC 1600 - Heavy Construction Other Than Building Const - Contractors

Construction Partners, Inc.

Fluor Corp

Granite Construction Inc

Great Lakes Dredge & Dock Corp

Jacobs Engineering Group Inc /De/

KBR, Inc.

Meadow Valley Corp

Orion Group Holdings Inc

Peter Kiewit Sons Inc /De/

Sterling Construction Co Inc

Williams Industrial Services Group Inc.

SIC 1623 - Water, Sewer, Pipeline, Comm And Power Line Construction

Aegion Corp

Dycom Industries Inc

Energy Services Of America Corp

Goldfield Corp

Infrasource Services Inc

Mastec Inc

MYR Group Inc.

Preformed Line Products Co

Primoris Services Corp

SIC 1700 - Construction Special Trade Contractors

ADM Endeavors, Inc.

Aduddell Industries Inc

Ameresco, Inc.

America Greener Technologies, Inc.

Argan Inc

Astro Aerospace Ltd.

Biopower Operations Corp

Brand Energy & Infrastructure Services, Inc

Cavico Corp

Concrete Pumping Holdings, Inc.

Diversified Global Holdings Group Inc.

Firemans Contractors, Inc.

Free Flow, Inc.

Fuquan Financial Co

Furmanite Corp

H/Cell Energy Corp

In Media Corp

Layne Christensen Co

Limbach Holdings, Inc.

Lime Energy Co.

Matrix Service Co

Peco Ii Inc

Powercomm Holdings Inc.

Real Goods Solar, Inc.

Reliant Holdings, Inc.

Solarcity Corp

Topbuild Corp

Us Home Systems Inc

Xstream Mobile Solutions Corp

SIC 3531 - Construction Machinery & Equip

Astec Industries Inc

ASV Holdings, Inc.

Caterpillar Inc

Columbus Mckinnon Corp

Douglas Dynamics, Inc

Gencor Industries Inc

JLG Industries Inc

Manitowoc Co Inc

US-BLH Bio-Engineering Int'l, Inc.

SIC 8711 - Engineering Services

Acorn Energy, Inc.

AECOM

Alion Science & Technology Corp

Aria International Holdings, Inc.

Biopharma Manufacturing Solutions Inc.

CH2M Hill Companies Ltd

Ecology & Environment Inc

Energy Edge Technologies Corp.

Engility Holdings, Inc.

Englobal Corp

Essex Corp

Galenfeha, Inc.

Hill International, Inc.

Infrastructure Developments Corp.

Mistras Group, Inc.

PBSJ Corp

Sotera Defense Solutions, Inc.

Tetra Tech Inc

TRC Companies Inc /De/

URS Corp /New/

Versar Inc

VSE Corp

Washington Group International Inc

Willdan Group, Inc.

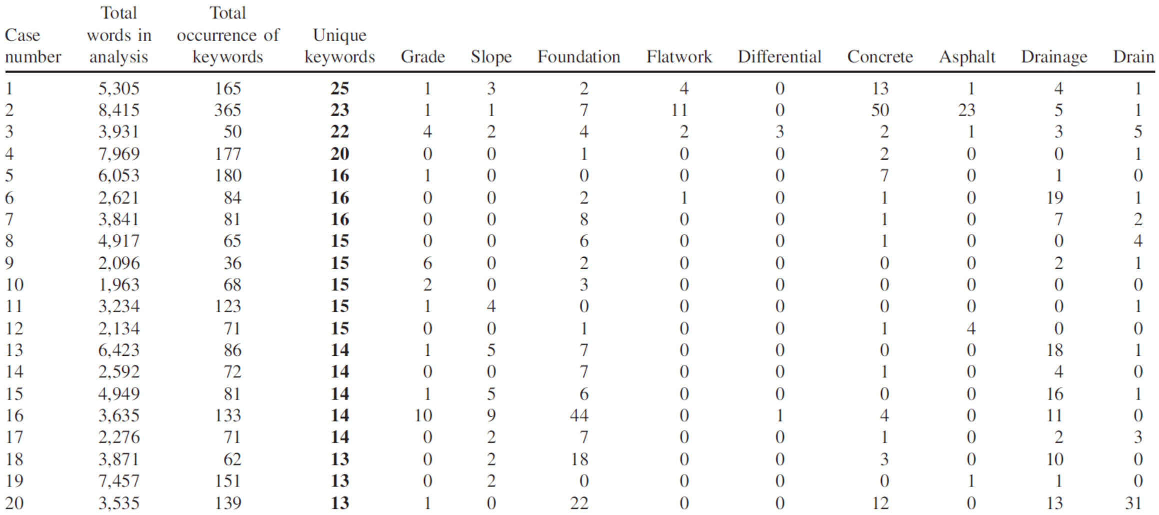
APPENDIX B. BRIEF SUMMARY OF APPLICATION OF TEXT MINING TO IDENTIFY PATTERNS IN CONSTRUCTION DEFECT LITIGATION CASES

(Note: Research conducted in collaboration with construction law experts at UC Denver)

The goal of this research was to build a text mining tool to identify patterns within the construction defect litigation cases by examining the language of public legal filings. It served as a pilot implementation of natural language processing and text mining to identify commonly occurring patterns and themes surrounding construction defect litigation. National legal database LexisNexis was used as a massive source of data comprising of summaries of thousands of past construction defect litigation cases. It led to a development of a pilot tool to crawl several hundred recent construction litigation cases and generate keywords and topics to facilitate content analysis procedure and perform a cursory exploration of the construction litigation landscape.

B1. Frequency Analysis of Important Keywords

The first analysis was a frequency analysis of pre-determined keywords. The intent of frequency analysis was to use keyword frequencies in the text as a proxy for issue relevance. The research sought to test if patterns in keyword frequencies could be identified as consequential. Based on earlier research works, 41 words with increased frequency in the plaintiff expert reports were identified as important by experts in the analysis of case files. A computer program was written using these keywords to read through a dataset of 1498 legal cases, after applying pre-processing and data cleaning techniques, to calculate the individual frequencies of the words across the cases. A sample of 20 cases with nine keywords of the highest frequency results is presented in Figure 22.



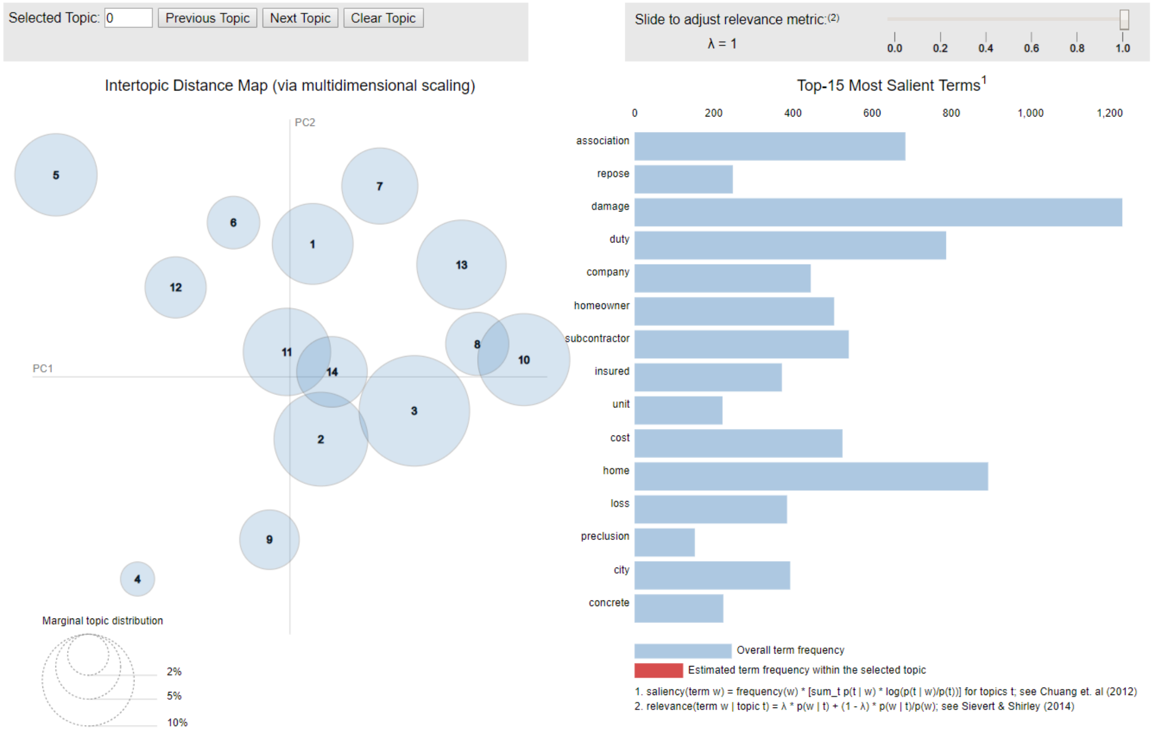
Case number	Total words in analysis	Total occurrence of keywords	Unique keywords	Grade	Slope	Foundation	Flatwork	Differential	Concrete	Asphalt	Drainage	Drain
1	5,305	165	**25**	1	3	2	4	0	13	1	4	1
2	8,415	365	**23**	1	1	7	11	0	50	23	5	1
3	3,931	50	**22**	4	2	4	2	3	2	1	3	5
4	7,969	177	**20**	0	0	1	0	0	2	0	0	1
5	6,053	180	**16**	1	0	0	0	0	7	0	1	0
6	2,621	84	**16**	0	0	2	1	0	1	0	19	1
7	3,841	81	**16**	0	0	8	0	0	1	0	7	2
8	4,917	65	**15**	0	0	6	0	0	1	0	0	4
9	2,096	36	**15**	6	0	2	0	0	0	0	2	1
10	1,963	68	**15**	2	0	3	0	0	0	0	0	0
11	3,234	123	**15**	1	4	0	0	0	0	0	0	1
12	2,134	71	**15**	0	0	1	0	0	1	4	0	0
13	6,423	86	**14**	1	5	7	0	0	0	0	18	1
14	2,592	72	**14**	0	0	7	0	0	1	0	4	0
15	4,949	81	**14**	1	5	6	0	0	0	0	16	1
16	3,635	133	**14**	10	9	44	0	1	4	0	11	0
17	2,276	71	**14**	0	2	7	0	0	1	0	2	3
18	3,871	62	**13**	0	2	18	0	0	3	0	10	0
19	7,457	151	**13**	0	2	0	0	0	0	1	1	0
20	3,535	139	**13**	1	0	22	0	0	12	0	13	31

Figure 22 - A sample of 20 cases with nine keywords of the highest frequency

Based on the results obtained, 366 cases containing 5 or more keywords were identified. The keywords ‘concrete,’ ‘window,’ ‘water,’ ‘roof,’ ‘foundation,’ and ‘structural,’ were the top six most commonly occurring words in this frequency analysis and each occurred more than 1,000 times in the data set with ‘water’ being the highest at 3,864 occurrences.

B2. Unsupervised Approach implementing Latent Dirichlet Allocation (LDA)

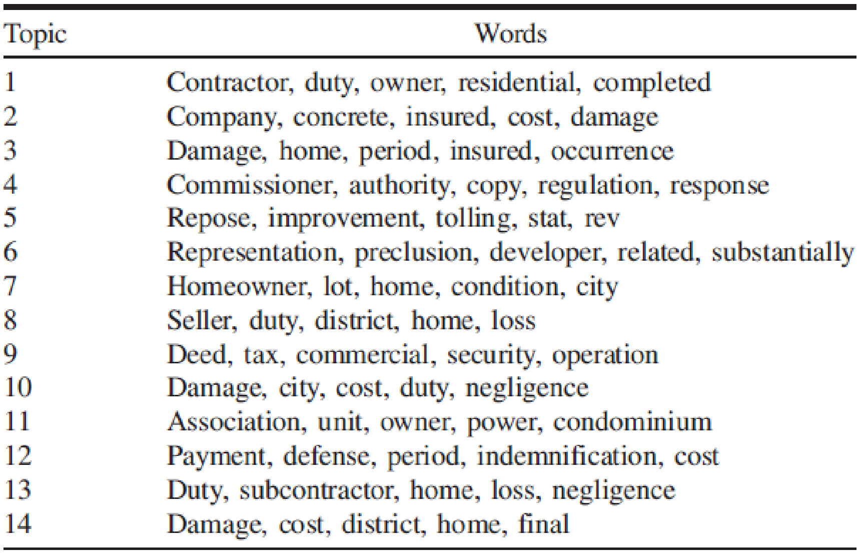
Unsupervised learning on text data involves applying algorithms on a dataset that has not been manually labeled, classified or categorized by the user. Instead, trends and patterns in the input data are found based on the inherent properties and characteristics of the data, rather than potential biases introduced by the user. The second analysis of this study, was the application of unsupervised approach on 59 cases, a subset of the 1498 cases that originated from the state of Colorado between the period 2000-2017. For analysis, the Latent Dirichlet Allocation (LDA) algorithm was implemented, which is commonly used in topic modelling for natural language processing applications. The goal was to develop a pilot implementation of LDA for a construction defect litigation dataset, and observe the topics/ themes that were produced by the algorithm in an unsupervised manner. A total of 14 topics were generated for the input data as shown in Figure 23.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 23 - Topics and the words associated with them visualized using the pyLDAvis interface

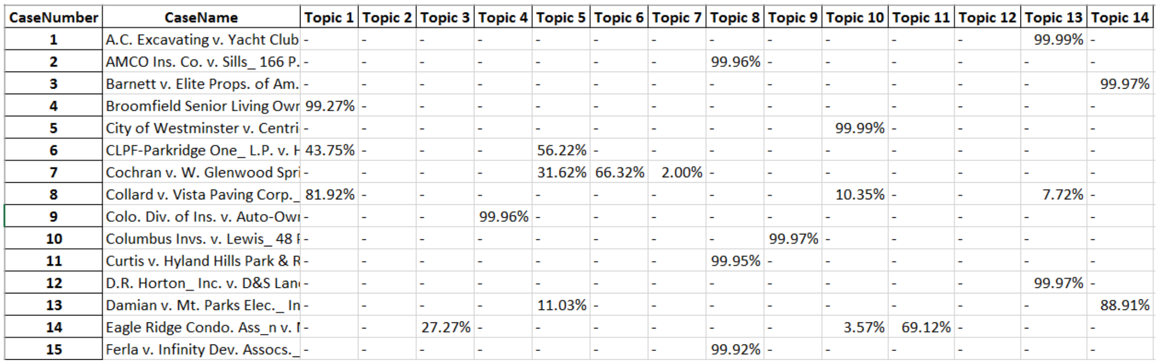
Each topic produced by LDA was a distribution of words with different weights and usually the top 5-10 words by weight are used to describe the overarching theme captured by the topic. Figure 24 presents the top 5 words obtained associated with each of these topics.



Topic	Words
1	Contractor, duty, owner, residential, completed
2	Company, concrete, insured, cost, damage
3	Damage, home, period, insured, occurrence
4	Commissioner, authority, copy, regulation, response
5	Repose, improvement, tolling, stat, rev
6	Representation, preclusion, developer, related, substantially
7	Homeowner, lot, home, condition, city
8	Seller, duty, district, home, loss
9	Deed, tax, commercial, security, operation
10	Damage, city, cost, duty, negligence
11	Association, unit, owner, power, condominium
12	Payment, defense, period, indemnification, cost
13	Duty, subcontractor, home, loss, negligence
14	Damage, cost, district, home, final

Figure 24 - Top 5 words representating each LDA Topic

Finally, individual legal cases were attributed to the topics obtained, as shown in Figure 25. Each case was allowed to be classified as one or more topics based on how closely they align with the theme captured in the underlying topics.



CaseNumber	CaseName	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
1	A.C. Excavating v. Yacht Club	-	-	-	-	-	-	-	-	-	-	-	-	99.99%	-
2	AMCO Ins. Co. v. Sills_ 166 P.	-	-	-	-	-	-	-	99.96%	-	-	-	-	-	-
3	Barnett v. Elite Props. of Am.	-	-	-	-	-	-	-	-	-	-	-	-	-	99.97%
4	Broomfield Senior Living Ow	99.27%	-	-	-	-	-	-	-	-	-	-	-	-	-
5	City of Westminster v. Centri	-	-	-	-	-	-	-	-	-	99.99%	-	-	-	-
6	CLPF-Parkridge One_ L.P. v.	43.75%	-	-	-	56.22%	-	-	-	-	-	-	-	-	-
7	Cochran v. W. Glenwood Spri	-	-	-	-	31.62%	66.32%	2.00%	-	-	-	-	-	-	-
8	Collard v. Vista Paving Corp._	81.92%	-	-	-	-	-	-	-	-	10.35%	-	-	7.72%	-
9	Colo. Div. of Ins. v. Auto-Ow	-	-	-	99.96%	-	-	-	-	-	-	-	-	-	-
10	Columbus Invs. v. Lewis_ 48	-	-	-	-	-	-	-	-	99.97%	-	-	-	-	-
11	Curtis v. Hyland Hills Park & R	-	-	-	-	-	-	-	99.95%	-	-	-	-	-	-
12	D.R. Horton_ Inc. v. D&S Lan	-	-	-	-	-	-	-	-	-	-	-	-	99.97%	-
13	Damian v. Mt. Parks Elec._ In	-	-	-	-	11.03%	-	-	-	-	-	-	-	-	88.91%
14	Eagle Ridge Condo. Ass_n v.	-	-	27.27%	-	-	-	-	-	-	3.57%	69.12%	-	-	-
15	Ferla v. Infinity Dev. Assocs._	-	-	-	-	-	-	-	99.92%	-	-	-	-	-	-

Figure 25 - Mapping of the cases to identified topics

REFERENCES

Adams, F.K. “Construction Contract Risk Management: A Study of Practices in the United Kingdom”, Cost Engineering Vol. 50/No. (2008).

Al-Bahar, J.F. and Crandall, K.C. “Systematic Risk Management Approach for Construction Projects”, J. Constr. Eng. Manage., (1990), 116(3): 533-546.

Alomari, K.A., Gambatese, J.A., Tymvios, N. “Risk Perception Comparison among Construction Safety Professionals: Delphi Perspective”, J. Constr. Eng. Manage., (2018), 144(12): 04018107.

Ashraf, R. “Scraping EDGAR with Python”, The Journal of Education for Business 92(1):1-7 · May (2017) DOI: 10.1080/08832323.2017.1323720.

Ashuri, B., Jallan, Y. and Lee, J.H. “Materials Quality Management for Alternative Project Delivery”, Georgia DOT Research Project Number RP 16-22, (2018).

Ashuri, B., Moradi, Arash., Baek, Minsoo., Kingsley, Gordon., Yehyun, Hannah., Zhang, Limao., Liang, Yuping and Bahrami, Shiva. “Risk Mitigation Strategies to Enhance the Delivery of Highway Projects”, Georgia DOT Research Project Number RP 16-40, (2018).

Bao, Y. and Datta, A. “Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures”, Journal of Management Science, Vol. 60, No. 6 (2014), pp. 1371–1391.

Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation. (n.d.). Retrieved November 25, 2019, from https://www.crummy.com/software/BeautifulSoup/bs4/doc/#.

Berendt B., “Text Mining for News and Blogs Analysis”, Encyclopedia of Machine Learning and Data Mining (2017). Springer, Boston, MA.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). "Latent Dirichlet Allocation", Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Bojanowski, P., E. Grave, A. Joulin, T. Mikolov (2016). “Enriching Word Vectors with Subword Information". arXiv:1607.04606.

Brogan, E., McConnell, W. and Clevenger, C.M. (2018) “Emerging Patterns in Construction Defect Litigation: Survey of Construction Cases”, J. Legal. Affairs and Dispute Resolution, 2018, 10(4): 03718003.

Caldas, C.H. and Soibelman, L. “Automating hierarchical document classification for construction management information systems”, Journal of Automation in Construction, 12 (4) (2003) 395–406.

Campbell, J.H., Chen, H., Dhaliwal, D.S., Lu, H.M., Steele, L.B. “The information content of mandatory risk factor disclosures in corporate filings”. Review of Accounting Studies (2014) 19:396–455.

Chin, M.V., Liu, Y., Moffitt, K. “Voluntary Disclosure Through the Ranking of Risk Factors in the 10-K”, (2018) https://ssrn.com/abstract=3142990.

Choudhary, A.K., Oluikpe, P.I., Harding J.A. and Carrillo, P.M. “The needs and benefits of Text Mining applications on Post-Project Reviews”, Journal of Computers in Industry 60 (2009) 728–740.

Carrillo, P., Harding, J. and Choudhary, A. “Knowledge discovery from post-project reviews”, Journal of Construction Management and Economics, (2011) 29:7, 713-723, DOI: 10.1080/01446193.2011.588953.

Comprehensive Search Page. (n.d.). Retrieved February 1, 2020, from https://www.sec.gov/search/search.htm.

Construction Cybercrime Is on the Rise | 2019-05-08 | Engineering News-Record. (n.d.). Retrieved January 20, 2020, from https://www.enr.com/articles/46832-construction-cybercrime-is-on-the-rise.

Creedy, G.D., Skitmore, M., Wong, J.K. “Evaluation of Risk Factors Leading to Cost Overrun in Delivery of Highway Construction Projects”, J. Constr. Eng. Manage., (2010), 136(5): 528-537.

Design-Build Moves from Alternative to Mainstream – DBIA. (n.d.). Retrieved January 20, 2020, from https://dbia.org/design-build-moves-from-alternative-to-mainstream/.

Directory listing of full-index/. (n.d.). Retrieved February 13, 2020, from https://www.sec.gov/Archives/edgar/full-index/.

Division of Corporation Finance SIC Code List. (n.d.). Retrieved February 1, 2020, from https://www.sec.gov/info/edgar/siccodes.htm.

Dumais, S.T. "Latent Semantic Analysis". Annual Review of Information Science and Technology (2005) 38: 188–230. doi:10.1002/aris.1440380105.

English · spaCy Models Documentation. (n.d.). Retrieved February 13, 2020, from https://spacy.io/models/en#en_core_web_lg.

Finishing Acquisition of Layne Christensen, Granite Construction Revamps Roles | 2018-06-20 | Engineering News-Record. (n.d.). Retrieved December 10, 2019, from https://www.enr.com/articles/44736-finishing-acquisition-of-layne-christensen-granite-construction-revamps-roles.

Francis, L. and Flynn, M., “Text Mining Handbook”, Casualty Actuarial Society E-Forum, (2010).

Gad, G.M., and Shane, J. “Culture-Risk-Trust Model for Dispute-Resolution Method Selection in International Construction Contracts”, Journal of Legal Affairs and Dispute Resolution, (2017), 9(4): 04517020.

Gensim: Topic modelling for humans. (n.d.). Retrieved November 25, 2019, from https://radimrehurek.com/gensim/.

Granite Construction Incorporated FORM 10-K. (n.d.). Retrieved February 13, 2020, from https://www.sec.gov/Archives/edgar/data/861459/000086145918000008/gva1231201710k.htm.

Hallowell, M.R., Molenaar, K.R. and Fortunato, B.R. “Enterprise Risk Management Strategies for State Departments of Transportation”, J. Manage. Eng., (2013), 29(2): 114-121.

Huang, K.W. and Li, Z. “A Multilabel Text Classification Algorithm for Labeling Risk Factors in SEC Form 10-K”, ACM Transactions on Management Information Systems, Vol. 2, No. 3, Article 18 (2011).

Jallan, Y., Brogan, E., Ashuri, A., Clevenger, C. “Application of Natural Language Processing and Text Mining to Identify Patterns in Construction Defect Litigation Cases”, Journal of Legal Affairs and Dispute Resolution, (2019), 11(4): 04519024.

Jarkas, A.M. and Haupt, T.C. "Major construction risk factors considered by general contractors in Qatar", Journal of Engineering, Design and Technology, Vol. 13 Issue: 1, pp.165-194, (2015) https://doi.org/10.1108/JEDT-03-2014-0012.

Le, T., and Jeong, H.D. “NLP-Based Approach to Semantic Classification of Heterogeneous Transportation Asset Data Terminology”, Journal of Computing in Civil Engineering, (2017), 31(6): 04017057.

Lee, J.H., Yi, J.S., and Son, J.W., “Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP” Journal of Computing in Civil Engineering (2019), 33(3): 04019003, DOI: 10.1061/(ASCE)CP.1943-5487.0000807.

Leonard, A. (2018). Corporate reputation risk in relation to the social media landscape (Doctoral dissertation).

Mahfouz, T., Kandil, A., Davlyatov, S. “Identification of latent legal knowledge in differing site condition (DSC) litigations”, Journal of Automation in Construction, 94 (2018) 104–111.

Marzouk, M. and Enaba, M. “Text analytics to analyze and monitor construction project contract and correspondence”, Journal of Automation in Construction, 98 (2019) 265–274.

Matplotlib: Python plotting — Matplotlib 3.1.3 documentation. (n.d.). Retrieved February 13, 2020, from https://matplotlib.org/.

Miihkinen, A. “The usefulness of firm risk disclosures under different firm riskiness, investor-interest, and market conditions: New evidence from Finland”, Journal of Advances in Accounting, incorporating Advances in International Accounting 29 (2013) 312–331.

Mikolov, Tomas; et al. (2013). “Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781.

Mirakur, Y. “Risk Disclosure in SEC Corporate Filings”, (2011). Wharton Research Scholars. 85. https://repository.upenn.edu/wharton_research_scholars/85.

Moon, S., Shin, Y., Hwang, B.G., Chi, S. “Document Management System using Text Mining for Information Acquisition of International Construction”, KSCE Journal of Civil Engineering (2018) 22(12):4791-4798.

Natural Language Toolkit — NLTK 3.4.5 documentation. (n.d.). Retrieved February 13, 2020, from https://www.nltk.org/.

Python-edgar · PyPI. (n.d.). Retrieved February 1, 2020, from https://pypi.org/project/python-edgar/.

Qady, M.A and Kandil, A. “Automatic clustering of construction project documents based on textual similarity”, Journal of Automation in Construction 42 (2014) 36–49.

Rajaraman, A. and Ullman, J.D. (2011). "Data Mining", Mining of Massive Datasets. pp. 1–17. doi:10.1017/CBO9781139058452.002. ISBN 978-1-139-05845-2.

Re — Regular Expression operations — Python 3.8.0 documentation. (n.d.). Retrieved November 25, 2019, from https://docs.Python.org/3/library/re.html.

Requests · PyPI. (n.d.). Retrieved February 1, 2020, from https://pypi.org/project/requests/.

Sarbanes-Oxley Act of 2002. (n.d.). Retrieved February 13, 2020, from https://www.congress.gov/107/plaws/publ204/PLAW-107publ204.pdf.

Seaborn: statistical data visualization — seaborn 0.10.0 documentation. (n.d.). Retrieved February 13, 2020, from https://seaborn.pydata.org/.

SEC.gov | About EDGAR. (n.d.). Retrieved November 25, 2019, from https://www.sec.gov/edgar/aboutedgar.htm.

SEC.gov | Form 10-K. (n.d.). Retrieved November 25, 2019, from https://www.sec.gov/fast-answers/answers-form10khtm.html.

SEC.gov | How to Read a 10-K. (n.d.). Retrieved November 25, 2019, from https://www.sec.gov/fast-answers/answersreada10khtm.html.

SEC.gov | Smaller Reporting Companies. (n.d.). Retrieved February 13, 2020, from https://www.sec.gov/smallbusiness/goingpublic/SRC.

SEC.gov | What We Do. (n.d.). Retrieved November 25, 2019, from https://www.sec.gov/Article/whatwedo.html.

Sievert, C. and Shirley, K.E. “LDAvis: A method for visualizing and interpreting topics”, Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces (2014), pages 63–70, Baltimore, Maryland, USA, June 27, 2014.

Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.

Spielholz, P., Davis, D., Griffith, J. “Physical risk factors and Controls for musculoskeletal disorders in construction trades”, Journal of Construction Engineering and Management, 132 (10) (2006), pp. 1059-1068.

Stemming and lemmatization. (n.d.). Retrieved February 2, 2020, from https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

Tixier, A., Hallowell, M., Rajagopalan, B., Bowman, D. “Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports”, Journal of Automation in Construction 62 (2016) 45–56.

Touran, A. “A Mathematical Structure for Modeling Uncertainty in Cost, Schedule, and Escalation Factor in a Portfolio of Projects”, Construction Research Congress (2014), ASCE, Reston, VA, 1743–1751.

Tran, D.Q., and Molenaar, K.R. “Risk-Based Project Delivery Selection Model for Highway Design and Construction”, J. Constr. Eng. Manage., (2015), 141(12): 04015041.

Two Teams Certify Costs for Boston Green Line Extension | 2017-11-01 | ENR. (n.d.). Retrieved December 10, 2019, from https://www.enr.com/articles/43334-two-teams-certify-costs-for-boston-green-line-extension.

United Airlines: Stock Drops Following Passenger Incident in Chicago | Fortune. (n.d.). Retrieved January 17, 2020, from https://fortune.com/2017/04/11/united-airlines-stock-drop/.

Wang, D., Dai, F., Ning, X., Dong, R.G., Wu, J.Z. “Assessing Work-Related Risk Factors on Low Back Disorders among Roofing Workers”, J. Constr. Eng. Manage., (2017), 143(7): 04017026.

Williams, T.P. and Betak, J.F. “Identifying Themes in Railroad Equipment Accidents Using Text Mining and Text Visualization”, International Conference on Transportation and Development (2016).

Williams, T.P. and Gong, J. “Predicting construction cost overruns using text mining, numerical data and ensemble classifiers”, Journal of Automation in Construction 43 (2014) 23–29.

WordNet | A Lexical Database for English. (n.d.). Retrieved February 13, 2020, from https://wordnet.princeton.edu/.

Yarmohammadi, S., Pourabolghasem, R., Shirazi, A. and Ashuri, B. “A Sequential Pattern Mining Approach to Extract Information from BIM Design Log Files”, (2016) 33rd International Symposium on Automation and Robotics in Construction.

Zhang, L. and Ashuri, B. “BIM log mining: Discovering social networks”, Journal of Automation in Construction, 91 (2018) 31-43.

Zhao, X., Hwang, B.G., Gao, Y. “Fuzzy Synthetic Evaluation Approach for Risk Assessment: a case of Singapore's Green Projects”, Journal of Cleaner Production 115 (2016) 203-213.

VITA

Yashovardhan Jallan was born in Golaghat, Assam, India. He completed his Bachelor’s degree in Civil Engineering along with a Master’s degree in Structural Engineering from Indian Institute of Technology (IIT) Kharagpur, India in May 2016. He began his Doctoral program in Civil Engineering in August 2016, under the advisement of Dr. Baabak Ashuri at the Georgia Institute of Technology, Atlanta, U.S.A. During his graduate studies at Georgia Tech, he completed two Master’s degrees with one in Quantitative and Computational Finance and the other one in Computational Science and Engineering. His research area was interdisciplinary in nature which combined knowledge of computational methods and finance applied to practical issues in the civil engineering and construction industry.