FRAMEWORK FOR AUTOMATED INFORMATION EXTRACTION FROM CONSTRUCTION CORRESPONDENCE

By

ADEEL REHMAN

(NUST2017MSCE&M0900000207006)

Master of Science

In

Construction Engineering and Management

Department of Construction Engineering and Management

School of Civil and Environmental Engineering (SCEE)

National University of Sciences and Technology (NUST),

Islamabad, Pakistan

(2021)

This is to certify that the

thesis titled

FRAMEWORK FOR AUTOMATED INFORMATION EXTRACTION FROM CONSTRUCTION CORRESPONDENCE

Submitted by

ADEEL REHMAN

(NUST2017MSCE&M0900000207006)

has been accepted towards the partial fulfillment
of the requirements for the degree of
Master of Science in Construction Engineering and Management

Dr. Muhammad Usman Hassan
Supervisor / Assistant Professor
Department of Construction Engineering and Management
School of Civil and Environmental Engineering (SCEE)
National University of Sciences and Technology (NUST)

THESIS ACCEPTANCE CERTIFICATE

It is certified that the final copy of MS thesis written by Adeel Rehman (Registration No. NUST2017MSCE&M0900000207006), of PG Wing – SCEE has been vetted by undersigned, found complete in all respects as per NUST Statutes / Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for the award of MS/MPhil degree. It is further certified that necessary amendments, as pointed out by GEC members of the scholar, have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: Dr. Muhammad Usman Hassan

Date: _____

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

DEDICATED

TO

MY LOVING PARENTS, BROTHERS AND FRIENDS

ACKNOWLEDGMENTS

I, Adeel Rehman, am thankful to Allah Almighty for giving me the strength to carry out the research work. Furthermore, I am obliged to my Thesis Supervisor, Dr. M. Usman Hassan, for his valuable guidance, time, and encouragement. I also owe acknowledgments to my parents' patience, prayers, and support which has helped me stay steadfast throughout my life, especially in this challenging endeavor.

I am highly indebted to Ms. Kinza Rubab, who has been a tremendous help in the execution and validation of this study. I am thankful to all the respondents who were seminal to completion of this research. I pay gratitude to the respected GEC members for their constant support and guidance. I am grateful to my friends, the esteemed faculty and administration of the Department of Construction Engineering and Management (CE&M) of National University of Sciences and Technology (NUST), Pakistan, for giving the much-needed technical inputs, assistance, and resources for the thesis work.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Content Analysis	CA
Manual Content Analysis	MCA
Automated Content Analysis	ACA
Text Mining	TM
Vector Space Model	VSM
Data Mining	DM
Naïve Baes	NB
Natural Language Processing	NLP
Machine Learning	ML
Deep Learning	DL
Comma Separated Values	CSV
Information Extraction	IE
Information Retrieval	IR
Artificial Intelligence	AI
Optical Character Recognition	OCR
Parts of Speech	POS
Phrase Structure Grammar	PSG
International Federation of Consulting Engineers	FIDIC
True Positives	TP
False Positive	FP
False Negative	FN
True Negative	TN

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Communication of information is seminal to the success of any project, particularly the construction industry. A huge amount of data is generated daily in a construction project, the bulk of which is available in textual form. All this data is traditionally analyzed manually which is a process marred with time delays, cost ineffectiveness, error-prone, etc. To automate the process of content analysis, Text Mining has been used extensively in unstructured texts, but it falls short of understanding human language. Natural Language Processing (NLP) is a new AI based approach which allows a computer to understand texts in a human-like manner. Rule-based NLP was chosen for better results as the area of application is specific. A framework for automated information extraction from construction correspondence was formed and a Rule based NLP system reflecting the framework was also created. The inputs for the system are the corpus, manually defined Information Extraction rules, ontology, and a standard letter format. The question that what is to be extracted from a letter was put to field experts and their opinion was reflected by making headers of a summary table. For validation, the system was fed with Sixty letters from an existing project, and results were verified by field experts. The metric used was the F-1 score which is a harmonic mean of recall and precision. The score obtained, after repeated tuning of rules, was above ninety-five percent, signifying that the framework can be implemented to automate correspondence content analysis in construction projects.

Chapter 1

INTRODUCTION

1.1 BACKGROUND

The communication of information is seminal to the success of any project (Hollings and Centre, 1999). It not only serves as a backbone to any organization by proving robustness but also forms the basis for the types of organizational networks (Wiesenfeld et al., 1999). Moreover, it is an integral factor of traditional and non-traditional project delivery methods (Konchar, Sanvido, and Members, 1998). Construction project communication instruments are primarily in textual form and may include, inter-alia, day-to-day correspondence, meeting minutes, interim reports, project dashboards, presentations, emails, and others (Gibson and Cohen, 2003; Stackman and Henderson, 2010). The traditional approach to managing information is the manual analysis and subsequent extraction of useful information from the produced correspondence documents.

1.2 PROBLEM STATEMENT

The traditional method of Manual content analysis (MCA) and information extraction (IE) is not only outdated but also leads to various discrepancies such as missing out on crucial information and misinterpretation, thus rendering an unreliable product (Graaf and Vossen, 2013). In addition, the delay caused due to low efficiency of a person leads to higher project cost which is accompanied by an excessive effort by people (Evans, Mcintosh, et al., 2007). In manual content analysis, the problems of loss of information, misinterpretation, and low efficiency are common (Zhang and El-gohary, 2016). The process is time-consuming and yields inconsistency in results (Ur-Rahman, 2017; Wang *et al.*, 2018; Jallan *et al.*, 2019; Lee, Yi, and Son, 2019; Salminen *et al.*, 2019). Difficulty in the analysis of massive data and its reduced reusability are also the issues faced when it comes to MCA (Tixier *et al.*, 2016; Lee, Yi and Son, 2019; Salminen *et al.*, 2019). The inherent problems of data duplication and slow update of crucial information pose a hurdle to effective summarization of content (Graaf and Vossen, 2013; Lin and Su, 2013).

Several studies have used advanced computer technologies to address the issue of MCA and automate a specific process to minimize the influence of analysis by a human being. Zhang and El Gohary have created a Rule-based Natural Language Programming system for automatic information extraction from Construction Regulatory Documents (Zhang and El-gohary, 2016). Tixier et al. have automatically extracted precursors for injuries from a repository of reports based on common keywords (Tixier *et al.*, 2016). Niu and Issa have identified impact factors of claims for construction litigation cases automatically using an ontology-based assessment (Niu and Issa, 2014). Construction Correspondence from a completed project has been analyzed to find hidden information through word correlation using an online text mining tool (Marzouk and Enaba, 2019). None of the above-mentioned studies have extracted information from correspondence data, thus leaving a gap for a new study. So the question arises whether we can efficiently extract useful information from construction correspondence automatically with minimal human input.

In brief, the traditional approach to information extraction from construction correspondence, i.e. manual content analysis, is a process filled with various problems, thus calling for an automated method to cater for such discrepancies.

1.3 RESEARCH OBJECTIVES

- To identify inefficiencies in information extraction from construction correspondence using traditional content analysis techniques.
- To develop a framework for efficient and accurate extraction of information from construction correspondence.
- To evaluate the developed Framework to measure its performance.

1.4 SIGNIFICANCE OF STUDY

The construction industry in today's day and age is characterized by fragmentation, disintegration, and persistent complexity in activities and processes (Alashwal, Rahman and Beksin, 2011). Moreover, the said attributes can throttle the progress of a project and its success by compromising the delivery of key project objectives. To alleviate this dire situation, augmenting interaction and cooperation

among various entities within the industry can increase productivity yielding increased efficiency (Hollings and Centre, 1999; Garbharran and Govender, 2012). In the modern information age, accurate and latest information access is seminal to the formation of cooperation mechanisms and communication tools.

Automation in content analysis and processing of construction correspondence can yield easier information management, timely response, effective information handling, and increased efficiency. In addition to these, it can also lead to lesser subjectivity, minimizing information loss and lesser effort. Nearly eighty percent of corporate information is available in textual format and can be subjected to text analytics and mining for information extraction (Ur-Rahman and Harding, 2012).The sheer amount of correspondence documents engendered daily on a construction project proves to be a significant hurdle to data analysis and further classification (Alsubaey, Asadi, and Makatsoris, 2015).

On a national level, Pakistan's government and private sector are way behind the world in terms of automation in general and text analytics in particular. The construction industry contributes around seven percent to the national income given by the Pakistan Bureau of Statistics (PBS). Automation in the processing of construction correspondence can prove to be advantageous to its overall financial yield. Automation of the processing of correspondence can lead to decreased costs, higher efficiency, timely dissemination, and better information handling (Martínez-Rojas, Marin, and Amparo Vila, 2012).

This dissertation shall provide a comprehensive literature review defining traditional methods of content analysis its problems and a technique to resolve these issues. The literature review is followed by research methodology, in which the framework and automated system is created. Results for a selected case study are presented to experts for validation and their input. Concluding remarks are provided in the end.

Chapter 2

LITERATURE REVIEW

2.1 GENERAL

Construction worldwide tends to be an extremely information-dependent industry in which the success of any project is largely dependent on fair access to, the effective management of, and comprehensive analysis of communication or correspondence data (Martínez-Rojas, Marin and Amparo Vila, 2012). The construction industry is uniquely positioned because it entails characteristics that differentiate it from all the other industries. For instance, a construction project takes place over an extended period of time. It may include on-site production, a large number of people are involved, and more importantly, the company staff is highly variable (Zwikael, 2009). Additionally, during the project lifecycle, an enormous amount of documents with relevant data are produced and exchanged (Caldas, Soibelman, and Han, 2002a)

The primary modes of information exchange prevalent in construction projects are textual documents. Some of the key elements are contract documents, field reports, emails, and change orders. Thus, based on the structure of these documents, effective information extraction and analysis of the contained content becomes a challenge (Alsubaey, Asadi, and Makatsoris, 2015).

2.2 CONTENT ANALYSIS AND INFORMATION EXTRACTION

Content analysis (CA) is a process of analyzing messages in any type of information in a systemic manner and retrieve information. It has been described as “a technique which lies at the crossroads of qualitative and quantitative methods” and a technique that “allows a quantitative analysis of seemingly qualitative data (Duncan, 1989).

The development of inferences by extracting required information about any topic of interest to the reader in any type of communication is simply content analysis. In other words, the application of a pre-defined classification scheme on

any raw data containing text, images, and illustrations forms the process of content analysis (Kondracki, Wellman, and Amundson, 2002). Whereas the content may range from simple words and phrases to full-fledge essays, topics, and even scholarly articles consisting of theories and concepts.

Information Extraction is simply reducing the data contained in a document or text to a tabular form and has been pondered for years. Its rudimentary execution for medical knowledge-based documents was carried out at New York University by Naomi Sager (Sager, 1988).

Once found, the data or information may be analyzed either quantitatively or qualitatively or both ways. In a quantitative analysis, message components can be counted to identify specific themes, relative emphasis on various topics, amount of space or time devoted to certain topics, and numerous other dimensions which may prove beneficial if seen in a holistic perspective. In comparison, Qualitative analysis can be applied to determine latent meanings of the data under consideration. CA can yield specific patterns and enable theorists to theorize based on their ability and knowledge from the uncovered inferred ideas in the text. (Berg, 2004).

2.3 TRADITIONAL METHOD OF CONTENT ANALYSIS MANUAL CONTENT ANALYSIS AND INFORMATION EXTRACTION

The traditional method of content analysis and information extraction used in the industry is highly reliant on manual input and management of construction correspondence. Manual Content analysis is defined as a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding through manual labor. Holsti (1969) offers a broad definition of any type of content analysis as "any technique for making inferences by objectively and systematically identifying specified characteristics of messages." (Stemler, 2001)

To put it in our context, the thorough perusal, sifting out of relevant information from an analysis of documents, communication logs, field reports, emails, and other correspondence documents by people can be referred to as manual content analysis.

Textual data is exchanged at an alarming rate in the construction industry and can easily be regarded as the primary mode of exchange of information. Based on the structure of documents such as field reports, contracts, variation orders, communication logs, etc., the effective management of information becomes a major issue (Alsubaey, Asadi, and Makatsoris, 2015). A manual approach of establishing connections among the information extracted from such documents and their analysis is highly impractical due to the enormous amount of data stored in these documents (Graaf and Vossen, 2013; Alsubaey, Asadi, and Makatsoris, 2015).

MCA is time intensive as reading a document and summarizing it in a tabulated form takes time (Kondracki, Wellman and Amundson, 2002). In fact, if calculated in terms of man-hours, the process is costly and engenders other related costs (Fan and Li, 2013). Knowledge extraction and retrieval of useful information from given structured or unstructured data is a complex process and requires expertise thus the act of CA is highly influenced by the skill level of a person (Al Qady and Kandil, 2013). Moreover, the extracted information is not always consistent and varies from person to person. In fact, understanding a particular text is subjective in nature, thus reducing the relevancy of the extracted information (Evans, Mcintosh, *et al.*, 2007).

In addition, the reusability of such information remains a challenge as there is no specific structure (Tixier *et al.*, 2016). Manual IE also poses a great hurdle when it comes to the analysis of massive data because the work efficiency may vary with time and person resultingly extracted output shall be error-prone (Zhang and El-gohary, 2016). Furthermore, MCA leads to the loss of crucial information and slower update of information (Martínez-Rojas, Marin and Amparo Vila, 2012; Lin, Su and Chen, 2014). Hence, the MCA and IE is a process marred with many problems that are listed in Table 1.

Table 1- Inefficiencies in Manual Content Analysis and Information Extraction

S.No.	Inefficiencies in Manual Content Analysis and Information Extraction	References
1	Time-Consuming in processing (Time efficiency)	(Kondracki, Wellman and Amundson, 2002; Fonseca and Jorge, 2003; Evans, McIntosh, *et al.*, 2007; Matthes and Kohring, 2008; Wang, 2008; Eastman *et al.*, 2009; ESMAEILI and Matthew, 2012; Al Qady and Kandil, 2013; Lin and Su, 2013; Al Qady and Kandil, 2014a; De Graaf and Van Der Vossen, 2013; Fan and Li, 2013; Niu and Issa, 2014; Qady and Kandil, 2014; Villanova, 2014; Abbaszadegan and Grau, 2015; Niknam and Karshenas, 2015; Alsubaey, Asadi and Makatsoris, 2015; Meer, 2016; Tixier *et al.*, 2016; Zhang, El-gohary and Asce, 2016; Goh and Ubeynarayana, 2017; Nedeljkovic and Kovašević, 2017; Ur-Rahman, 2017; Wang *et al.*, 2018; Jallan *et al.*, 2019; Lee, Yi and Son, 2019; Salminen *et al.*, 2019)
2	Costly	(Evans, McIntosh, *et al.*, 2007; Eastman *et al.*, 2009; ESMAEILI and Matthew, 2012; Martínez-Rojas, Marin and Amparo Vila, 2012; W. Der Yu and Hsu, 2013a; Fan and Li, 2013; Villanova, 2014; Abbaszadegan and Grau, 2015; Niknam and Karshenas, 2015; Zhang, El-gohary and Asce, 2016; Meer, 2016; Tixier *et al.*, 2016; Ur-Rahman, 2017; Goh and Ubeynarayana, 2017)
3	The complexity of text retrieval	(Fonseca and Jorge, 2003; Evans, McIntosh, *et al.*, 2007; Matthes and Kohring, 2008; Martínez-Rojas, Marin and Amparo Vila, 2012; De Graaf and Van Der Vossen, 2013; Fan and Li, 2013; Al Qady and Kandil, 2014a; Alsubaey, Asadi and Makatsoris, 2015; Niknam and Karshenas, 2015; Tixier *et al.*, 2016; Zhang, El-gohary and Asce, 2016; Nedeljkovic and Kovašević, 2017)

4	Lack of Structure and Consistency of data	(Kondracki, Wellman and Amundson, 2002; Fonseca and Jorge, 2003; Evans, McIntosh, *et al.*, 2007; Martínez-Rojas, Marin and Amparo Vila, 2012; Fan and Li, 2013; Alsubaey, Asadi and Makatsoris, 2015; Tixier *et al.*, 2016; Goh and Ubeynarayana, 2017; Nedeljkovic and Kovašević, 2017; Salminen *et al.*, 2019)
5	Difficulty in Massive Data Analysis	(Caldas, Soibelman and Han, 2002b; Evans, McIntosh, *et al.*, 2007; Matthes and Kohring, 2008; ESMAEILI and Matthew, 2012; Martínez-Rojas, Marin and Amparo Vila, 2012; Al Qady and Kandil, 2013, 2014b; W. Der Yu and Hsu, 2013b; De Graaf and Van Der Vossen, 2013; Niu and Issa, 2014; Villanova, 2014; Niknam and Karshenas, 2015; Alsubaey, Asadi and Makatsoris, 2015; Salminen *et al.*, 2019; Jallan *et al.*, 2019; Lee, Yi and Son, 2019)
6	Impact of Personal Skill level	(Fonseca and Jorge, 2003; Matthes and Kohring, 2008; Mani, Feniosky and Savarese, 2009; Martínez-Rojas, Marin and Amparo Vila, 2012; W. Der Yu and Hsu, 2013b; Al Qady and Kandil, 2014a; Niu and Issa, 2014; Alsubaey, Asadi and Makatsoris, 2015; Jallan *et al.*, 2019)
7	Reduced Reusability (Scope of application)	(Caldas, Soibelman and Han, 2002b; Evans, McIntosh, *et al.*, 2007; Martínez-Rojas, Marin and Amparo Vila, 2012; W. Der Yu and Hsu, 2013b; Meer, 2016; Tixier *et al.*, 2016; Nedeljkovic and Kovašević, 2017)
8	Error-Prone in Large data	(Kondracki, Wellman and Amundson, 2002; Fonseca and Jorge, 2003; Evans, McIntosh, *et al.*, 2007; Mani, Feniosky and Savarese, 2009; Lin and Su, 2013; Lin, Su and Chen, 2014; Villanova, 2014; Zhang and El-gohary, 2016; Wang *et al.*, 2018; Jallan *et al.*, 2019; Salminen *et al.*, 2019)
9	Data Re-entry	(Wang, 2008; Mani, Feniosky and Savarese, 2009; Martínez-Rojas, Marin and Amparo Vila, 2012; Lin and Su, 2013; Lin, Su and Chen, 2014)

10	Reduced Work Efficiency	(Matthes and Kohring, 2008; Martínez-Rojas, Marin and Amparo Vila, 2012; De Graaf and Van Der Vossen, 2013; Fan and Li, 2013; Abbaszadegan and Grau, 2015; Zhang, El-gohary and Asce, 2016; Nedeljkovic and Kovašević, 2017; Wang *et al.*, 2018)
11	Ineffective Visual Representation	(Martínez-Rojas, Marin and Amparo Vila, 2012)(Mani, Feniosky and Savarese, 2009; Nedeljkovic and Kovašević, 2017; Wang *et al.*, 2018)
12	Data Redundancy/ Duplication of Data	(Wang, 2008; De Graaf and Van Der Vossen, 2013; Niknam and Karshenas, 2015)
13	Relevancy (Reduced objectivity)	(Kondracki, Wellman and Amundson, 2002; Evans, McIntosh, *et al.*, 2007; Al Qady and Kandil, 2014a; Villanova, 2014; Wang *et al.*, 2018; Jallan *et al.*, 2019)
14	Data loss (Storage Problems)	(Martínez-Rojas, Marin and Amparo Vila, 2012)
15	Inadequate Data Maintenance	(Martínez-Rojas, Marin and Amparo Vila, 2012; De Graaf and Van Der Vossen, 2013; Zhang, El-gohary and Asce, 2016)
16	Slow Update of Information	(Lin and Su, 2013)

2.4 AUTOMATED CONTENT ANALYSIS AND INFORMATION EXTRACTION

Automated Content Analysis is the usage of various techniques and algorithms to extract meaningful patterns and associations from large textual documents by computers. Due to the well-acknowledged and infinite potential of computers and advanced technology to render valuable enhancements in various industries, the construction industry has also undertaken various initiatives to support different activities in the construction project cycle (Martínez-Rojas, Marin and Amparo Vila, 2012). The field of ICT that deals with automated content analysis are text mining. With the application of specialized techniques, text mining can automate the process of information extraction.

2.4.1 TEXT MINING

The knowledge-intensive process in which an analyst works on a collection of documents by using a specific set of analysis tools is known as Text Mining (Grossman and Frieder, 2004). It is the process of examining large collections of documents to discover new information or help answer specific research questions.

Text Mining is also known as Knowledge Discovery from Text (KDT) or Document Information Mining. It is a process to identify the latent or inferred data and knowledge contained in documents (Sullivan, 2001). Techniques such as Data Mining (DM), Information Retrieval (IR), Computational Linguistics, Natural Language Processing (NLP), and Knowledge Representation are usually employed.

Text Mining vs Data Mining

Text Mining is sometimes erroneously referred to as data mining. It is a relatively new field of research that provides a software-based computational structure that can analyze a given data to find patterns. This new domain is now being widely used in various operations such as text-to-speech synthesis, email spam detection, etc. (Alsubaey, Asadi and Makatsoris, 2015).

The terms data mining and text mining are wrongly assumed to be the same thing where therein exists a subtle difference between the two which renders them quite disparate in reality. The critical difference between traditional Data Mining and the newly developed science of text mining is that structured data is the primary focus for the former. In contrast, the latter analyzes semi-structured or even nonstructured data (W. Der Yu and Hsu, 2013a). Dorre et al. addressed two challenges while applying text mining techniques and processes: (1) the manual approach for characteristic analysis of massive document-based data is highly inefficient, and (2) the definition of key attributes to categorize a large textual data collectively is not simple (Dörre, Jochen; Gerst, Peterl; Seiffert, 1999). A supplementary data preparation activity is needed to apply Text Mining in comparison to Data Mining.

In other words, data mining involves the application of statistical analysis and machine learning techniques allowing for the examining of predominant patterns in a database. The utilization of data mining in the processing of unstructured texts

which are in huge amounts is very limited due to its characteristics (Yoon and Park, 2004).

As an answer to the demand for analysis of unstructured documents, text mining emerged as a novel technique that has been used to perform the extraction of relevant data in the said domain. In short, text mining assigns a collection of tags on each text document, and then discovery operations are performed on these tags. These labels/tags are usually assigned to certain keywords in a document. The text mining algorithms then extract required information in any document by featuring these keywords. Recently, the interest in text mining has increased considerably, thus leading to its extensive use in knowledge management (Feldman *et al.*, 1998). Text mining is an emerging field and area of research and helps uncover information in plain sight (Marzouk and Enaba, 2019). The retrieval of information from a text by computer is possible through rule-based or machine learning algorithms (Brill and Mooney, 1997).

Previous research using Text mining

The industries which generate vast amounts of data are actively making use of information computer technologies to effectively categorize, classify, store and then retrieve specific data to increase its usability and also automate the process for time and cost-effectiveness (Evans, Mcintosh, *et al.*, 2007; Al Qady and Kandil, 2014b). Engineering projects in general and construction in specific are filled with instances where text mining and data analytics has been used to classify, retrieve and categorize data effectively.

Alsubaey has used text mining to classify and categorize construction project data and then formulate an Early Warning System to identify early warnings of failure in a construction project by feeding specific keywords and training the model with a massive amount of data. The system uses the ability of text mining technique named Naïve Bayes Classifier to learn from specified keywords' frequencies in certain texts and classify them as a specific early warning category (Alsubaey, Asadi and Makatsoris, 2015).

A Bayes Classification system determines the class of a document by checking a specific keyword's frequency and probability in a test set. Early warnings were identified by feeding the system a corpus of critical project management

documents such as meeting minutes. The technique was evaluated by taking feedback from experts and also unseen critical minutes of the meeting.

Fan and Li developed a text mining operator using a vector space model and extract similar cases for alternate dispute resolution in construction accidents. The model utilizes indexes or keywords to represent a file. These keywords or concepts are terms that can aptly reflect the data in the document and describe it. Words like claims, liabilities and specific injury types are cases-in-point. The cases were retrieved from a purpose-built library from Westlaw, and the test data set was given in simple text form. The system was judged based on the index or keywords whether they reflected the information in the case file by numerical metric Euclidean distance (Fan and Li, 2013).

A study by Caldas has categorized Construction project documents into pre-set categories such as general, schedule, demolition-civil, landscape-site, structures, interior finishes, HVAC, etc. by vector space model. The model represents terms per document and their specific frequency, thus effectively matching them to a certain category of documents making it easier to handle massive amounts of data and also enhance reusability (Caldas, Soibelman and Han, 2002a).

Wen-der applied a content-based text mining technique to retrieve Computer-aided design (CAD) drawings by extracting textual content from these files. Files were assigned with specific tags for better retrieval results. The technique used is similarity matching by employing a vector model which used numerical value to assign any cad file a similarity index, thus returning the closest match for any search word (W. Yu and Hsu, 2013).

Whereas, Fonseca has proposed a different method for retrieval of cad drawings by classifying, indexing, and finally retrieving technical drawings from a huge database by spatial relationships and elements of visual nature. In essence, graphical matching rather than textual matching was used. The metric of K- nearest neighbor was used to evaluate the System (Fonseca and Jorge, 2003).

Mohammed Al Qady has laid down a comparison of supervised and unsupervised text classifiers. The study has listed down the limitations of supervised classifiers and tried to explore the possibility of categorizing construction project documentation in pre-defined categories by applying Machine Learning text mining

techniques. The writers have come to the revelation that an amalgam of both supervised and unsupervised systems would fit the purpose of categorizing such documents. To evaluate the System, Al Qady has used the F-1 score as a measure, a harmonic mean of Recall and Precision (Al Qady and Kandil, 2014b).

Michael Evans et al. have provided an overview and general assessment of the machine learning techniques of text mining used for classification in the field of LAW to improve empirical legal research (Evans, Mcintosh, *et al.*, 2007). The writers have contended that the legal texts are not only lengthy but also complex in understanding thus acting as a challenge to experts to main consistency while coding such complicated documents especially when it comes to comparing various cases. The study has revealed that the Word scores method developed by Laver (2003) has proven effective as it assigns specific keywords scores and then computes their frequency in any document. Further, classification is effectively done by employing the Naïve Bayes model. In political fields, the given methods are proving to be helpful.

Salminen et al. have used machine learning to tag online content across various platforms and in varied forms, thus providing a valuable addition to content marketing efficiency (Salminen *et al.*, 2019). The study compares various machine learning techniques for the classification of multi-label content such as Random Forest, K-Nearest Neighbor, and neural network analysis by feeding the system created a considerable amount of online news articles for auto-tagging and further classification. F-1 score has used a measure to evaluate the performance of aforementioned techniques-based models, and Neural network analysis was revealed to return the best results.

Meer has suggested that automated content analysis can be used to effectively communicate during crises as it avoids the cost and time lost in the analysis of massive data sets of crisis communication documentation manually (Meer, 2016). The research proposes different methods that can be used to help analyze large amounts of documents.

2.4.2 NATURAL LANGUAGE PROCESSING (NLP)

After a thorough study of the literature above, it is evident that text mining can handle computational tasks very adeptly but understanding human language remains a challenge for traditional text mining clustering and classification techniques. As construction correspondence contains information in a natural language format, its direct extraction by computational devices was not effectively and accurately possible until the advent of modern text mining techniques using artificial intelligence such as Natural Language Processing (NLP) (Brill and Mooney, 1997). NLP is essentially the convergence of artificial intelligence, linguistics, and Information Computer Technology as illustrated in Figure 1. Its ultimate goal is to achieve an understanding of natural language as human beings (Liddy, 2001). It is a text mining technique that enables machines i.e. computers to process any text containing natural language in a human-like manner. (Cherpas, 1992).

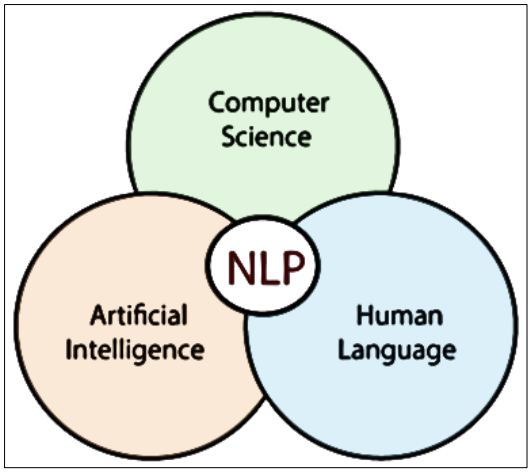

Figure 1- Natural Language Processing

It can help significantly in correctly identifying and extracting relevant information from documents by having its roots in a subject-specific ontology, i.e. a database of the vocabulary of the subject. Thus it not only processes a given text based on the syntax but also on semantics leading to much better results in terms of comprehension and information extraction (Zhang and El-gohary, 2012).

Information extraction (IE) is a subsidiary field of NLP that seeks to extract the required information from a textual document to populate pre-defined data templates. Information Extraction can be based on syntactic features, i.e.

grammatical analysis and semantic features, i.e. narrative analysis of the text (Zhang and El-gohary, 2016).

Some of the major uses of NLP include among others are machine translation, recognition of speech, and automated content analysis (Manning and Schütze, 1999). The application of Automated content analysis is progressively being adopted in different industries. The reason for this is the necessity to understand and effectively make use of burgeoning digitized data (Bai, 2011). The construction industry is loaded with electronic information which may be structured or unstructured. Construction projects, even smaller sized, have a plethora of documents in the shape of computer-aided drawings, specifications, inventory management, process control, scheduling, cost estimating, and other documentation (Soibelman *et al.*, 2008).

However, the formation of a structure that understands a text of natural language and its further validation is problematic. A natural language is an amalgam of words list known as lexicon; rules of their structure known as grammar that renders the meaning of words by putting them in their specified positions in a sentence (Manning and Schütze, 1999). Modelling grammatical rules has been a major issue in the past while analyzing natural languages (Hindle, 1989). In addition, a word may have different meanings in different contexts, leading to confusion incorrectly analyzing the writer's intended purpose by a computer. Modern methods based on statistics and machine learning procedures are used to tackle the issues mentioned above. The methods used for analyzing texts include clustering procedures e.g. k-means, and classification procedures e.g. support vector machines, naïve Bayes, and k-nearest neighbors (Manning and Schütze, 1999; Grimmer and Stewart, 2013; Alsubaey, Asadi and Makatsoris, 2015; Tixier *et al.*, 2016).

Rule-based NLP vs Machine Learning NLP

Two types of approaches are widely used in NLP: rule-based and machine learning approaches. The rule-based approach’s essence lies in manual coded guidelines which need to be revised repeatedly to effectively analyze any natural language text. On the other hand, a machine learning (ML)-based approach makes use of ML algorithms that instruct models which process a given text by feeding it massive amounts of related data for it to learn (Tierney, 2012).

As such a vast supply of data is required for ML-based NLP and the coding process takes up a lot of time and resources, so a manually coded rule-based approach could be the answer to this study.

According to Sage and Lavie, the use of rule-based NLP yields better results in terms of accuracy as it brings an element of human intelligence and data-related expertise (Sagae and Lavie, 2003). Wang thinks that the use of statistical methods i.e. ML or DL for classification, aims at shallow comprehension and broad viewpoint. In contrast, manually defined rules are relatively better at yielding deeper understanding within a particular area (Wang et al., 2002). Rule-based NLP can therefore handle sentence-level tasks, such as parsing and extraction very well thus making it a better fit for query analysis such as text extraction under specific headers. In addition, machine learning algorithms are relatively obscure in comparison to manually crafted rules and dictionaries as these can be easily updated (Breiman, 2001; Barbella et al., 2009).

NLP uses a set of fundamental text processing techniques in a certain order for achieving element recognition, relationship identification, and in turn extraction. The techniques are Tokenization, Sentence Splitting, Parts of Speech (POS) tagging, etc. (Cunningham *et al.*, 2011).

NLP in Construction Industry

Zhang and El-Gohary have used rule-based NLP in the automation of checking compliance for regulatory documents of construction. One chapter of the International Building Code (IBC) was used as a reference for compliance checking. The study employed pattern matching information extraction rules and in some cases conflict resolution rules as well. The NLP tool was created on a free software named GATE and the rules were coded in java. Validation of the NLP tool created was achieved by calculating the F-1 score against a gold standard formulated by experts.

The score turned out to be above ninety-four percent. (Zhang and El-gohary, 2016). The Framework suggested by them is given in Figure 2.

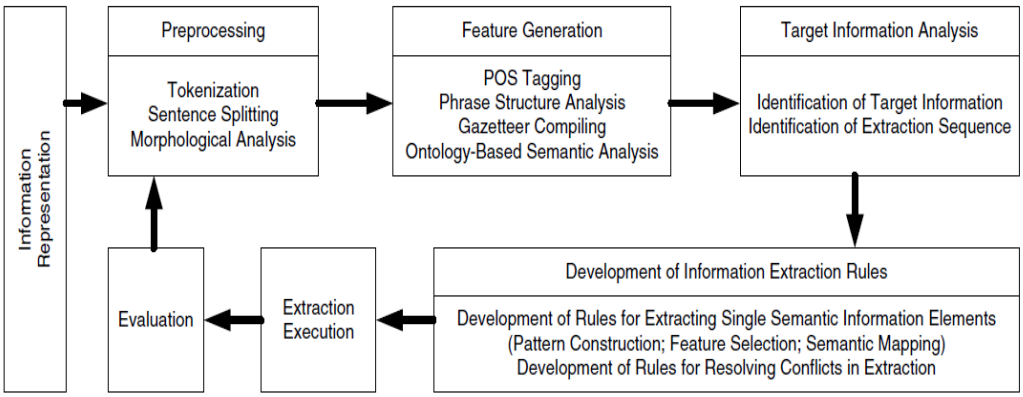

Figure 2- Zhang and El-Gohary's Rule based NLP Framework.

Tixier ET. all has used rule-based NLP for the extraction of precursors from unstructured injury reports, thus organizing such reports in structured formats. (Tixier *et al.*, 2016). The study used R programming language to code the rules manually and formed a keyword dictionary; ontology to analyze the content of the injury reports effectively. The premise Tixier based the analysis on is that any injury can be uniquely and adequately defined by a set of specific construction site characteristics. These traits named precursors by the author in any injury were used as keywords to search for and retrieve the closely matching incidents. The methodology adopted is given in Figure 3.

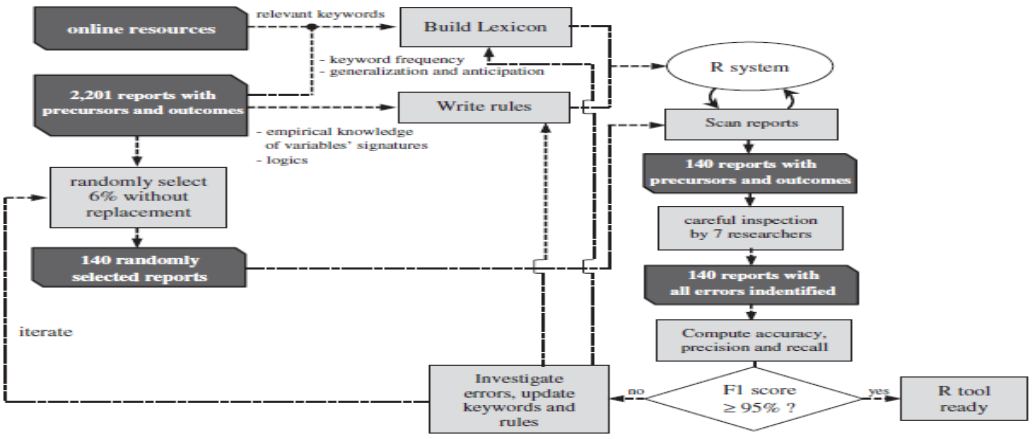

Figure 3- Tixier's Framework for Injury Reports Precursor Extraction

Niu and Issa have employed Rule-based Natural Language Processing to automate the impact factors understanding in a construction litigation case related to

claims (Niu and Issa, 2014). The paper has proposed a set of software that can be used together to achieve automation in claim factor identification. The example taken to elaborate the Framework is that of Differing Site Conditions Claims for which it has suggested collecting precedent claim cases from a professional legal database to build an ontology and then develop ontology-aware Java mapping rules in GATE software.

After a thorough study of the literature, it is not wrong to say that Natural Language Processing techniques are more than capable to handle any human-generated texts. As for the question of which specific technique to use rule base or ML-based approach following facts pushed the study towards a rule-based approach. First and foremost, (Chiticariu, Li and Reiss, 2013) have found that in private or public industry small and large companies rule-based NLP is preferred as they are more inclined towards higher qualitative results in specific domains rather than a wide area of application. They have named companies such as IBM, SAP, and Microsoft being adherents of the rule-based approach. Secondly, due to the unavailability of massive amounts of data and lack of time and resources Machine, learning-based NLP was not a favorable choice. Finally, the level of quality in results required was high so a Rule-Based approach to NLP making use of manually coded rules for information extraction was decided upon for the study. The studies provided above have all made use of Rule-based NLP.

Chapter 3

RESEARCH METHODOLOGY

This chapter describes the procedure for achieving the objectives outlined in chapter 1. The research is designed in compliance with the detailed research process including identification of critical issues in manual content analysis and development of Framework from literature.

3.1 RESEARCH DESIGN:

After the preliminary research in which the objectives were defined by finding out the gap within existing literature, the research followed three major stages in which stage 1 was the identification of inefficiencies and outlining critical constraints of Manual Content Analysis (MCA) as shown in Figure 4. In Stage 2 based on the study of various methods for automation in information extraction, a framework is developed having its roots in Rule-based NLP and then a system is created to test the proposed Framework. Stage 3 shows that the results are evaluated by validating them from field experts.

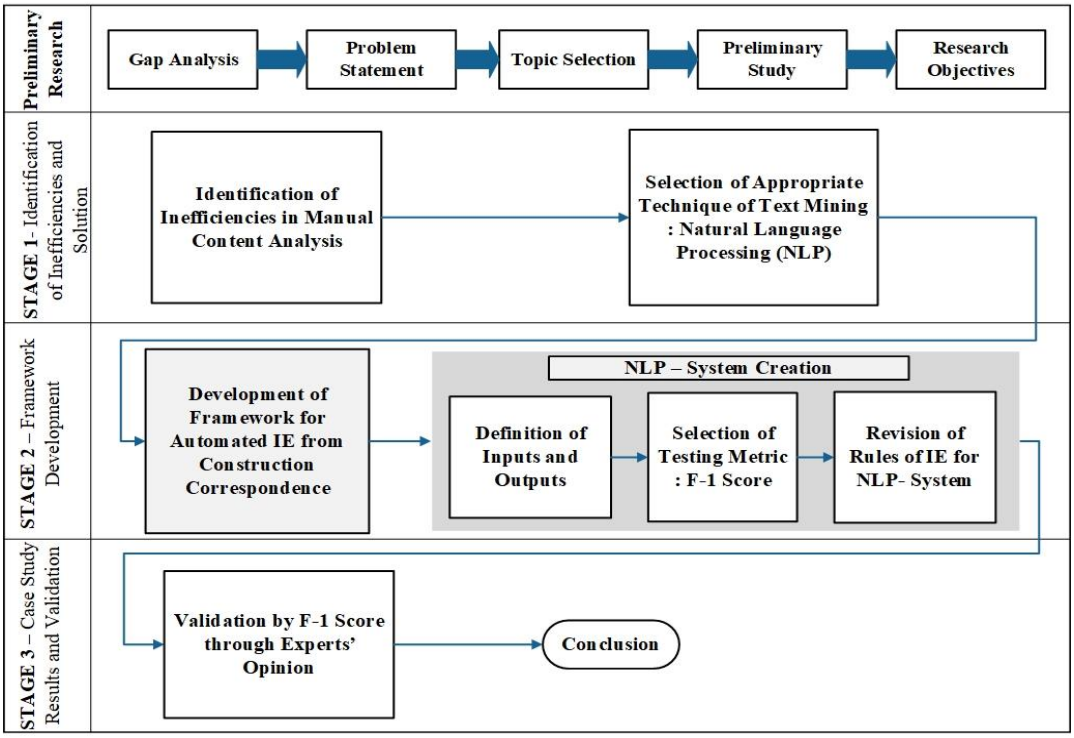

Figure 4-Research Methodology Steps

3.2 STAGE 1 - IDENTIFICATION OF INEFFICIENCIES in MCA and IE

Now that the domain was established a detailed literature review was carried out to rank the inefficiencies associated with MCA and IE according to literature. The basis for the literature score lies in both the qualitative aspect and quantitative aspect. For instance, how many research articles have acknowledged an inefficiency defined the quantitative aspect and how much influence it has laid on the said issue outlined the inefficiency's qualitative aspect. The inefficiencies with references have been explained in the literature review under the header of traditional method of content analysis. The literature score is given in Table 2.

Table 2- Literature Score and Ranking of Inefficiencies in MCA and IE

Rank Literature	**Inefficiencies in Manual Content Analysis and Information Extraction**	**Literature Score**
1	Time Consuming in processing (Time efficiency)	0.875
2	Costly	0.4375
3	Complexity of text retrieval	0.375
4	Lack of Structure and Consistency of data	0.3125
5	Difficulty in Massive Data Analysis	0.3
6	Impact of Personal Skill level	0.28125
7	Reduced Reusability (Scope of application)	0.21875
8	Error Prone in Large data	0.20625
9	Data Re-entry	0.15625
10	Reduced Work Efficiency	0.15
11	Ineffective Visual Representation	0.125
12	Data Redundancy/ Duplication of Data	0.05625
13	Relevancy (Reduced objectivity)	0.0375
14	Data loss (Storage Problems)	0.01875
15	Inadequate Data Maintenance	0.00625
16	Slow Update of Information	0.00625

Furthermore, to validate our findings in the literature, a preliminary field survey from the construction industry was conducted through which the inefficiencies were ranked instead of field scores only where the respondents were asked to evaluate how much efficiency would affect the effective extraction of information from a given text. Likert scale from 1 to 5 was used where 5 denoted very high influence and 1 denoted no influence at all. The results of the field survey are given in Table 3.

Table 3- Field Score and Ranking of Inefficiencies in MCA and IE

Rank	**Inefficiencies in Manual Content Analysis**	**Field Score**
1	Slow Update of Information	0.893333333
2	Time Consuming in processing (Time efficiency)	0.88
3	Error Prone in Large data	0.846666667
4	Reduced Work Efficiency	0.84
5	Costly	0.833333333
6	Difficulty in Massive Data Analysis	0.833333333
7	Relevancy (Reduced objectivity)	0.826666667
8	Lack of Structure and Consistency of data	0.806666667
9	Impact of Personal Skill level	0.786666667
10	Inadequate Data Maintenance	0.773333333
11	Complexity of text retrieval	0.766666667
12	Data loss (Storage Problems)	0.753333333
13	Data Re-entry	0.74
14	Reduced Reusability (Scope of application)	0.72
15	Data Redundancy/ Duplication of Data	0.72
16	Ineffective Visual Representation	0.666666667

The distribution of respondents, their respective experience, and organization type is given in Figure 5.

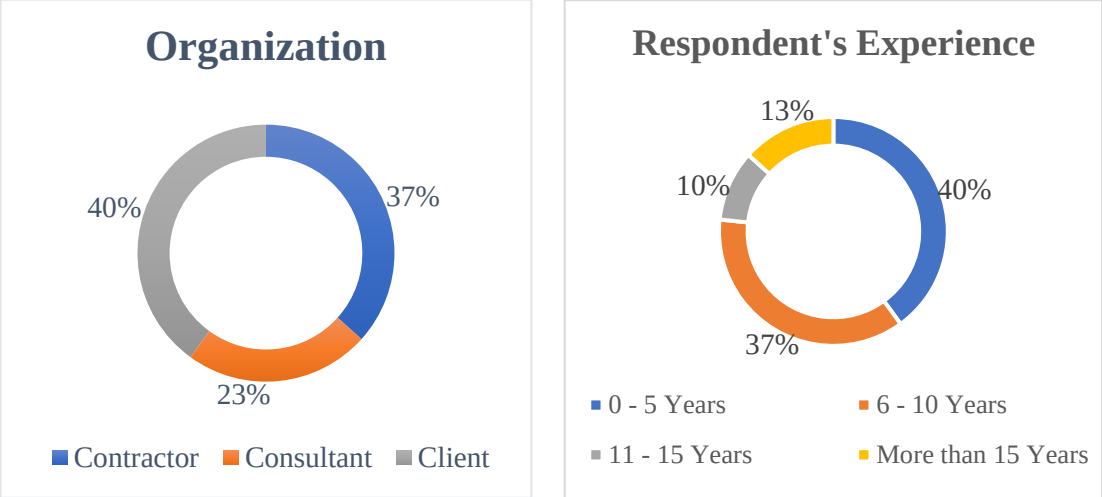

Figure 5- Organizational and Experience-based division of Respondents.

To evaluate and rank these inefficiencies based on field and literature, a Sixty Forty Ratio analysis was taken to include the influence of field and literature. Time-consuming, costly in nature, difficulty in massive data analysis, the complexity of text retrieval, and lack of structure were the critical issues of manual content analysis, with values ranging from 0.6 to 0.878, identified in the study. The inefficiencies with their final ranking and respective scores has been given in Table 4.

Table 4- Sixty-Forty ratio of Field and Literature and Final Ranking of Issues in Manual Content Analysis

Final Rank	**Inefficiencies in Manual Content Analysis and Information Extraction**	**Literature Score**	**Field Score**	**60-40 Analysis**
1	Time Consuming in processing (Time efficiency)	0.875	0.88	0.878
2	Costly	0.4375	0.833333333	0.675
3	Difficulty in Massive Data Analysis	0.3	0.833333333	0.62
4	Complexity of text retrieval	0.375	0.766666667	0.61
5	Lack of Structure and Consistency of data	0.3125	0.806666667	0.609

6	Error Prone in Large data	0.20625	0.846666667	0.5905
7	Impact of Personal Skill level	0.28125	0.786666667	0.5845
8	Reduced Work Efficiency	0.15	0.84	0.564
9	Slow Update of Information	0.00625	0.893333333	0.5385
10	Reduced Reusability (Scope of application)	0.21875	0.72	0.5195
11	Relevancy (Reduced objectivity)	0.0375	0.826666667	0.511
12	Data Re-entry	0.15625	0.74	0.5065
13	Inadequate Data Maintenance	0.00625	0.773333333	0.4665
14	Data loss (Storage Problems)	0.01875	0.753333333	0.4595
15	Data Redundancy/ Duplication of Data	0.05625	0.72	0.4545
16	Ineffective Visual Representation	0.125	0.666666667	0.45

3.3 STAGE 2 - DEVELOPMENT OF FRAMEWORK

As concluded from the previous chapter, a rule-based approach to NLP is more suitable for analyzing letters that are human-generated texts. A framework involving the application of rule-based natural language processing to extract information from construction correspondence is constructed by studying literature regarding different rule-based NLP techniques. The elements needed for framework development were identified. To start with, a corpus is needed which is essentially a collection of text on which a linguistic analysis can be applied. In addition, a standard letter pattern for better extraction of the required information is essential. Moreover, an ontology, i.e. subject specific vocabulary, which is inclusive enough to fully describe the corpus is developed. The specification of output dictates the formation of rules and working of the system as a whole. After a framework for automated IE was proposed as shown in Figure 6, its validation was done by creating a Rule-Based NLP system which was tested and tuned to achieve results from actual construction industry letters. The results of the developed system were evaluated by field professionals to achieve a score on a suitable metric.

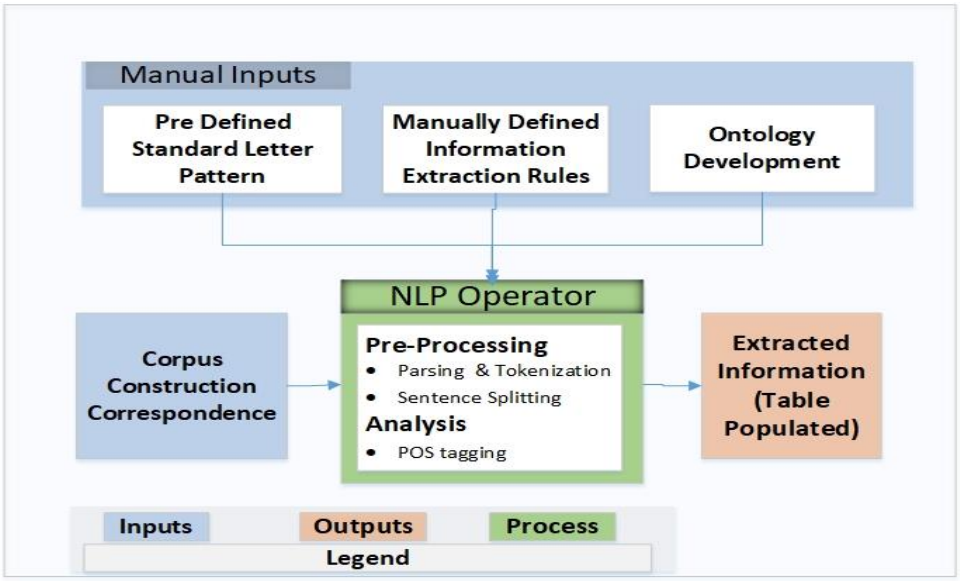

Figure 6-Rule-Based NLP Framework for IE from Construction Correspondence

3.3.1 NLP – System Input and Output Definition

Now that a conceptual framework was developed, an NLP - System was to be created on its basis. The task at hand was to define inputs and outputs required for the development of an NLP – System. A stepwise methodology was used which started with the study of letters from various organizations, defining components of a letter and in the end yielding corpus and ontology. The steps are given as a schematic diagram in Figure 7.

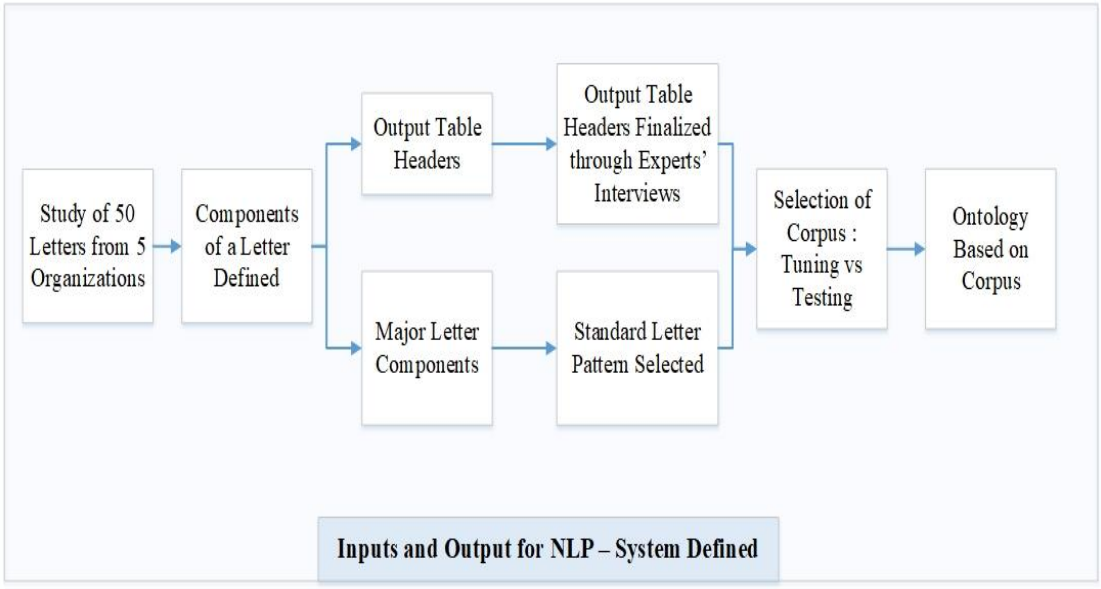

Figure 7- Inputs and Output for NLP- System Defined

Pre-Defined Standard Letter Pattern

Study of above fifty Letters from five reputed organizations, whether consultants, clients, or contractors, was conducted to better understand what constitutes a letter. The letter content parts common to all the organizations are listed below.

- Letter Date
- Letter No.
- Subject Heading
- Receiver details
- Letter no. showing who the letter was sent to and who has sent it
- Letter Replied to
- Sender Details
- Enclosure
- Copy to
- Subject may include the project name

There were certain unique components to different organizations as well. The type of organizations and the attributes unique to them have been given in Table 5.

Table 5 - Components Unique to organizations

S.No.	**Organization**	**Organization Type**	**Letter Content Parts**
1	Company A	Consultant	• Project details are given above the subject • References are listed down under a separate heading
2	Company B	Consultant	• Project Details are given under the Subject Heading • The subhead for the specific issue
3	Company C	Contractor	• Project Name and details under a Separate heading • The subject has been given separately

The selected standard letter pattern like any other corporate letter contains information of the sender and receiver i.e., name, post, address, and contact. Other than that, the date and specific letter-number used by the organization is given on top followed by sender details and then the project specifics and a subject heading. As a letter can address more than one reference letter but is chiefly a reply to the main letter so a heading for references is added. Following the list of references the body of the letter contains the information to be conveyed by the sender to the recipient and those they have identified in a heading below that the letter is "copy to". The annexures or extra documents enclosed have also been mentioned in the letter format used. In the end, all those officials or offices the correspondence is sent as a copy have been mentioned under the heading "copy to".

The format used by the consultant group taken as a standard letter pattern is given in Figure 8.

Ref: MG1/SCG/005/026 **(Letter ref. No.)** **(Date)**13 Jan, 2021

Receiver Details
John Stuart
CEO/ General Manager/ Project Director
Liaison Office, XYZ Development Company.
708 - WAPDA House, Lahore.
Telephone: +92-52-6920153172

SAMPLE PROJECT – Contract MG1-CG2 (Project Details)

Subject: Replies to the Comments on Sample Report

Ref: 1) DBDC Letter no. DBDC/W-10.12/3631-32 dated Nov 17, 2020 **(Letters Referred to)**

Dear Sir/ Madame,

(Body of the Letter)

This is with reference to your letter at Ref. 1) above through which comments of the Project Office and various formations of the company on Sample Report of the subject Project were conveyed to us.

Yours Sincerely,

(Sender Details)
Mr. ABC
Project Manager / Engineer's Representative
Sample Consultants Group

Encl: Replies to the Comments on Sample Report (26 Page)

Copy to: Site office

Figure 8-Standard letter pattern used.

Output Summary Table and Headers Defined

Each letter is scanned for extracting information using domain-specific keywords and the extracted phrase is parsed into a single record (an excel file) to store as one record per letter. The NLP system also computes counts of key indicators such as total letters, addressed, pending, total letters with clauses, costs, deadlines, delays, and enclosures. The headers for this record sheet were decided by conducting semi structures interviews of Six field professionals where they were asked to read a set of letter/s and then suggest necessary headers that could be used to summarize the letter content (Guest, Bunce and Johnson, 2006). The questions asked from the interviewees were:

- What information do you deem as important in the above letters?
- If we were to tabulate the data contained in these letters, what headers will you suggest?
- What additional headings do they recommend?

After getting their views on the above two questions, table headings suggested by the author with the help of the supervisor were shown. The organization type and the experience distribution of interviewees are shown in Figure 9.

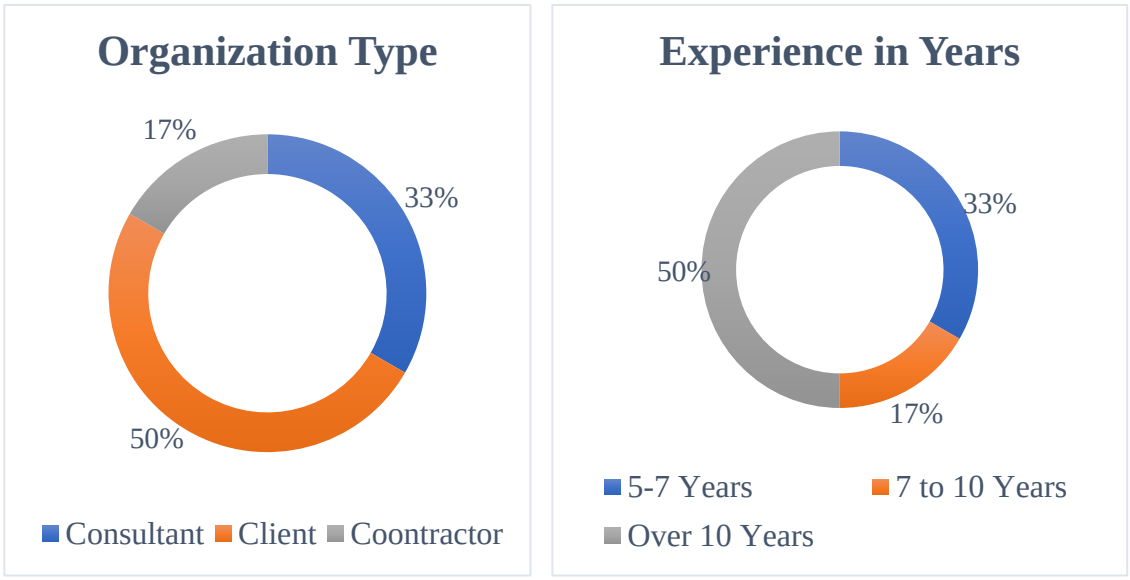

Figure 9- Details of interviewed Field Experts

The finalized headers for the summary sheet have been listed down.

- Letter Ref No.
- Date of Letter
- Reference Letter replied to
- All References
- Sender Name
- Designation of Sender
- Organization of sender
- Receiver Name
- Designation of Receiver
- Organization of Receiver
- Project Details
- Subject
- Purpose
- Clauses of Contract Involved
- Time Frame (if applied)
- Deadline date
- Days From Deadline
- Cost (if applied)
- Enclosed / Annex
- Copy to
- Status of Reply

Corpus of Construction Correspondence

Corpus is any collection of texts used for linguistic analysis. It may be a body of spoken or written content. For our study, a set of seventy letters from an existing project with their content largely based on management and contracts is taken as the corpus of construction correspondence. The study used actual letters relating to a dam project and consisted mainly of those sent by the consultant group of Six companies whose names have been kept confidential. Most of the letters were addressed to either the client or the two separate contractors. The corpus was divided in a ratio of 10 to 60 where ten letters were used for iterative tuning of the system and sixty were used as a case study to test the system.

Development of Ontology

The ontology is essentially a subject-specific vocabulary which helps in better understanding and recognition of various words by the computer. For out

study, the ontology was chiefly based on the corpus to point the NLP operator towards the desired output by provided keywords. As mentioned above the corpus was contractual letters so the FIDIC (International Federation of Consulting Engineers) Glossary for contracts was also kept as a repository to help in information extraction (FIDIC Glossary).

Manual Defined IE Rules

For information extraction, a list of headers based on the content of letters was selected, which could adequately summarize the data contained in a letter. Here the use of a standard letter format played a key role in identifying most of the data heads and subsequently the rules used for extraction as those were directed to extract an output matching to that header i.e. letter date, letter number, sender details, project specifics, subject, etc. Unlike the constant headings such as sender details or subject, etc., some of the headers may not return results if not present in the letter. For instance, the headers deadline date cost mentioned, time frame, clauses of the contract, etc. do not warrant the result in every letter.

3.3.2 Testing Metric and Validation Method

To evaluate and validate the effectiveness of an NLP-based system, a comprehensive approach is needed. As some domain-specific keywords in the ontology occur occasionally, we cannot use accuracy as a metric, as predicting no keywords most of the time will yield high accuracy. Therefore, a system of measurement that considers both multiple labels and the frequency of specified keywords is needed. The *F1 score*, which is the harmonic mean of two other metrics, Precision and Recall, is a suitable measure and has been used by various other studies as well (Wallach *et al.*, 2009; W. Der Yu and Hsu, 2013b; Al Qady and Kandil, 2014b; Tixier *et al.*, 2016; Zhang and El-gohary, 2016; Salminen *et al.*, 2019).

The F1 score may range from 0 to 1, such that 1 is the ideal case result. The Following equation, Eq. (1) shows how this metric is calculated.

$$F_1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad \text{Eq. (1)}$$

Where:

$$Precision = \frac{TP}{TP + FP} \quad \text{and} \quad Recall = \frac{TP}{TP + FN}$$

Precision is the measure of how well the developed NLP system avoids extracting the wrong output for a certain heading in the summary sheet. It is the ratio of true positives (TP) to the sum of true positives and false positives (FP) where TP is the instance where output is identified correctly, and FP is the one where output given is not correct.

On the other hand, *Recall* is a measure of how well the NLP system extracts outputs from the letters under a certain header of the summary sheet. It is calculated by dividing true positives by the sum of true positives and false negatives (FN). It should be noted that the last option, true negatives (TN), occurs when the system has not returned an output that is not present in a letter document. True negatives were not included. The harmonic mean of these two measurements yields the F1 score; thus, it considers both how well the system prevents the detection of incorrect outputs and how well it extracts the desired outputs.

3.3.3 NLP System Creation and Tunning

The whole system was created in an Anaconda distribution-based integrated development environment (IDE) named Spyder notebook, and Python was used as a language for coding. The schematic diagram for the creation and tuning of the NLP-System is given in Figure 10. The tuning corpus was separate from the testing corpus as mentioned above. It was a set of ten letters from the selected Dam project. The metric used was the F-1 score which is essentially a harmonic mean of Recall and Precision.

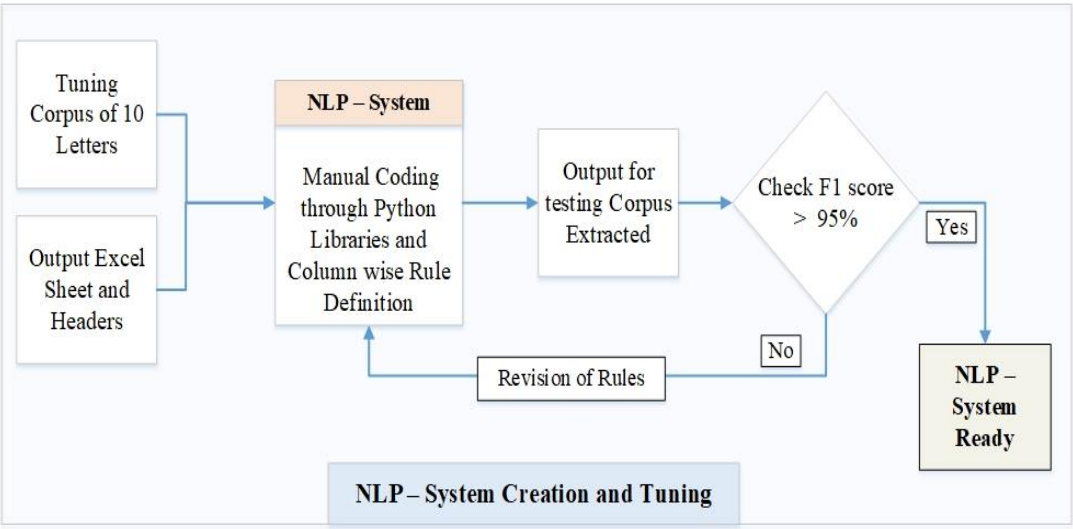

Figure 10- Creation of NLP System and its Tuning for Results

The system was used to extract information from the selected tunning corpus and then it was checked by calculating F-1 score. If the score was below 95 % which was selected as a threshold for the system performance the rules for extraction under those specific headers were refined to achieve better results in terms of extraction. The iterative revision of rules of IE used for each header of the output sheet was done to better extract information. For instance, if the system has not extracted sub clauses, then the keywords and rules were refined for the said column which were used to search for specific clauses and differentiate them from the clauses, articles, and provisions.

A *.bat extension file was created to easily execute the code for a layman to run as it can be used by pressing one button on the keyboard. The *.bat file has the details of the input directory for the file to process, output excels sheet, and processed folder, thus making the whole operator NLP-System portable. The operator automatically moves the processed files to a different folder after successfully processing the said files. The libraries used in Python and the interface of the SPYDER notebook can be seen in Figure 11.

The stored keywords are grouped into categories of interest in a predetermined format and visualized for further analysis and get insights from the data stored in the letters. The details of the NLP operator can be seen in figure 11 showing the NLP parsing engine and the algorithm for data retrieval. The NLP system also computes counts of key indicators such as total letters, addressed, pending, total letters with clauses, costs, deadlines, delays, and enclosures.

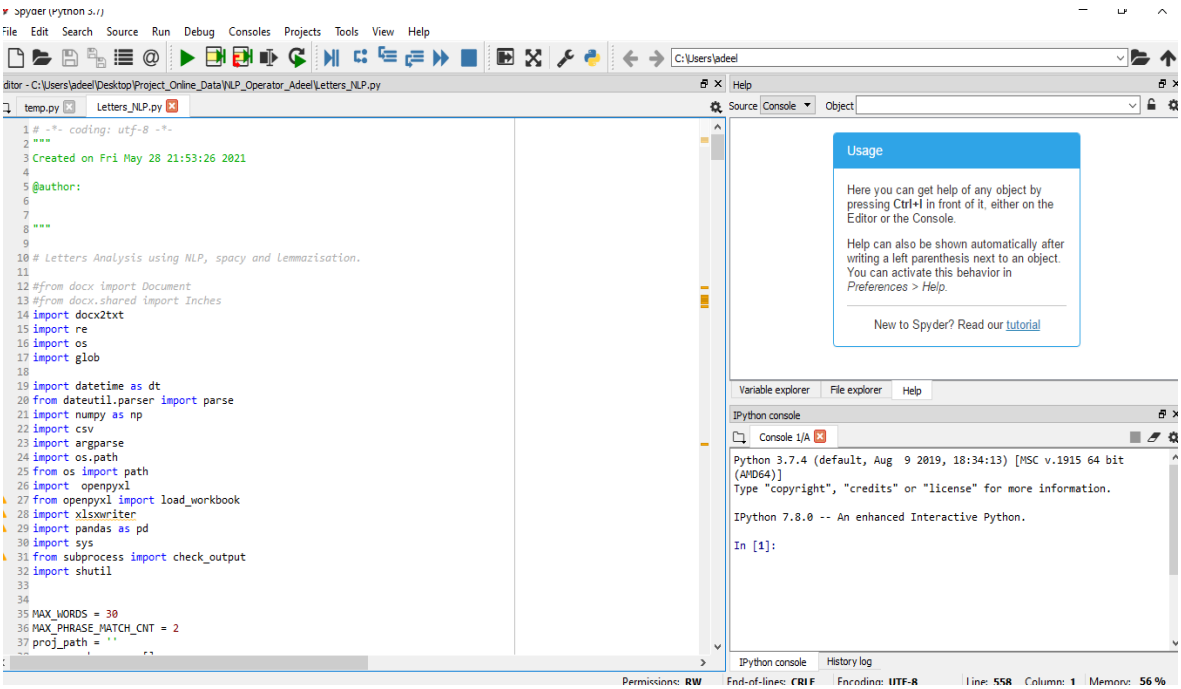

Figure 11- SPYDER Anaconda Interface

NLP System Algorithm

The algorithm for the NLP-System is shown in Figure 12. As shown, it gets a list of documents in word file format as input in an input folder named "data_to_process". These documents with the help of the doc2text library are converted to a text file and then the text is parsed into lines. Every line is then individually searched for a list of specific keywords. If the result is positive that is a match for a keyword is found, the line is tokenized into words and then pos-tagging is performed to get an idea of what function these words are performing. Using parsing and extracting features the exact phrase or sentence is extracted to fill a certain cell under the specific column. Then next line is picked up by the system to perform the same operation. After all keywords for all lines have been scanned then the system moves the document to the processed folder. A new document from the queue is selected and all the same, operations are performed for this file as well. Moreover, these documents are tokenized and parsed individually to retrieve domain-specific keywords and store them. The system writes one record for every letter processed successfully into an excel file, also updates the summary of processing the letters into a visualization sheet of an excel file. When there is an error

processing a letter, the letter is discarded and moved to the error path. The error file statistics or records are not used to compute any record or statistics.

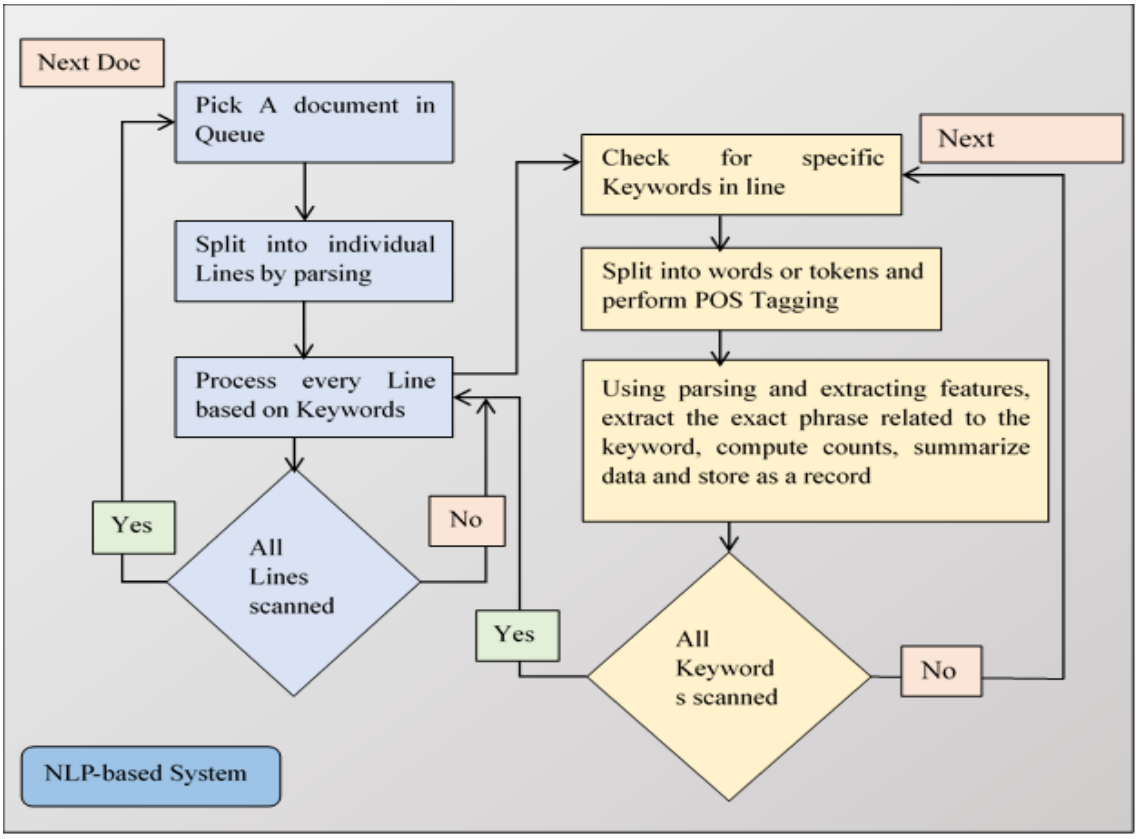

Figure 12-Algorithm for NLP System detailing Extraction sequence

In essence, a thorough analysis of literature and field led to the determination of critical constraints inherent to the process of MCA and IE thus signifying that an automated approach is needed. By studying literature, a subfield of text mining capable to understand human-generated text named the rule-based NLP approach was decided upon to automate the process of extraction of relevant information. A standard letter pattern was decided upon by perusal of different organization's correspondence and letter patterns. The headers of an output table were decided by conducting semi-structured interviews. All the inputs required to form the NLP system were finalized. Afterward, the NLP system was developed in Python, and rules are tuned and tested on a set of ten letters of the standard format but with different information to keep them separate as a test set. For validation, a set of sixty letters were fed to the NLP-System and their output was further assessed by experts.

Chapter 4

VALIDATION AND RESULTS

This chapter encompasses the validation method used, output of the NLP system, validation of the Framework by experts' review, and finally the resulting F-1 score.

4.1 VALIDATION BY EXPERTS

A set of ten test letters apart from the main corpus, collectively containing information that delivered a considerable output for every header, were taken to test and tune the rules for better results and achieving an F1 score of above Ninety Percent. This target in itself was quite an obstacle to surmount leading to repeated tweaking and tuning of rules. Afterward, the system was used to extract information from a total of Sixty (60) letters and populate an excel sheet i.e., the summary table.

As the corpus was chiefly contract and project management-based documents so the summary sheet populated by the NLP tool was evaluated by taking input from field experts having the relevant experience. Each of the experts was given a set of ten to twelve letters to read and then evaluate the information automatically extracted by the tool. They were asked to identify the correct outputs i.e., true positives (TP), wrong outputs i.e., false positives (FP), and missing outputs i.e. false negatives (FN) for these letters under every header which were verified by the writer to avoid any missing entries.

In instances where output provided by the system was correct but inadequate incapacity to properly justify the header, the expert was required to score the correctness of the output by giving it a score from 0 to 1; where 0 signified missing information making it a false negative and 1 being a true positive. Any score in between was divided as TP and FN such that a score of 0.8 would mean 0.8 is added to the TP and 0.2 to FN. Similarly, in some cases, extra information is extracted, so it was scored from 0 to 1 and dividing the rating between TP being the needed part and FP being the extra information as above. To better understand, cases with the extra output and inadequate extracted information are given in Figure 13 with their headers.

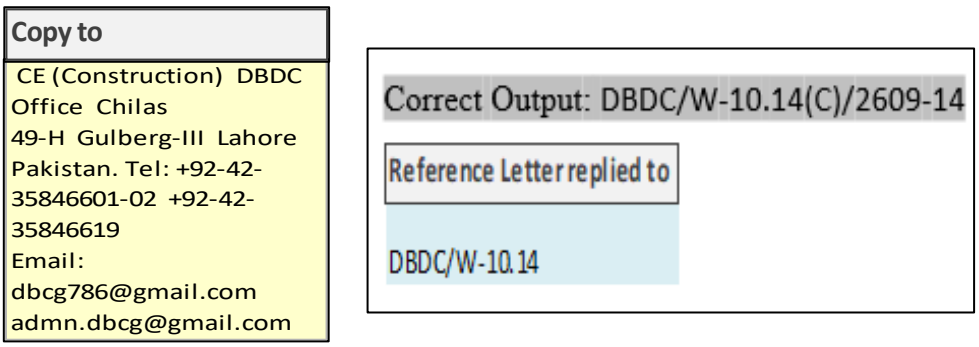

Figure 13- Examples of Extra output (left) and Inadequate output.

4.2 RESULTS & DISCUSSION

The result obtained from the NLP system tool is an excel sheet. The F-1 score was calculated by the values of TP, FP, and FN given by field experts for the NLP system populated excel sheet. The metric amounted to be 95.31 % signifying that rule-based NLP is effective in information extraction when specific outputs in a small area of application are required. The recall came out to be 95.24 % and precision was 95.38 % signifying that not only correct outputs were extracted in most of the cases, but also negative outputs were avoided as well thus achieving high values of precision and recall.

In this case study, those specific outputs were defined by the headers specified by the field experts and the specific area of application was contract and management-based letters. The scores achieved are considerably higher than those achieved through machine learning-based statistical models available in the literature. For example, Verma et al. analyzed tweets and categorized them into five classes by the employing Naïve Bayes Classification method and achieved an accuracy of 80% (Verma, Sudha; Vieweg, 2011). Bai has tried to predict the sentiment of consumers by automated mining of opinion from online texts containing movie reviews and news and achieved a range of accuracy from 66% and 88.9% (Bai, 2011). Automatic content analysis was performed by Grimmer and Stewart in the analysis of political texts by employing a random forest classifier and attained 65% accuracy with a 75% recall rate (Grimmer and Stewart, 2013).

Talking specifically about the construction industry Yu and Hsu were able to achieve an F-1 Score of 73% while retrieving Computer-aided drawings automatically due to low precision (W. Yu and Hsu, 2013). Al Qady's algorithms

for automatic unsupervised classification of construction projects documentation were done while achieving an F-1 score of 84.4% (Al Qady and Kandil, 2014a). (Tixier *et al.*, 2020) was able to obtain a very high F-1 Score of 96% while extracting precursors from a database of unstructured injury reports using Rule-based NLP. The same results have been compared in Figure 14 with all of the entries from construction industry except for Verma's classification of tweets (2011).

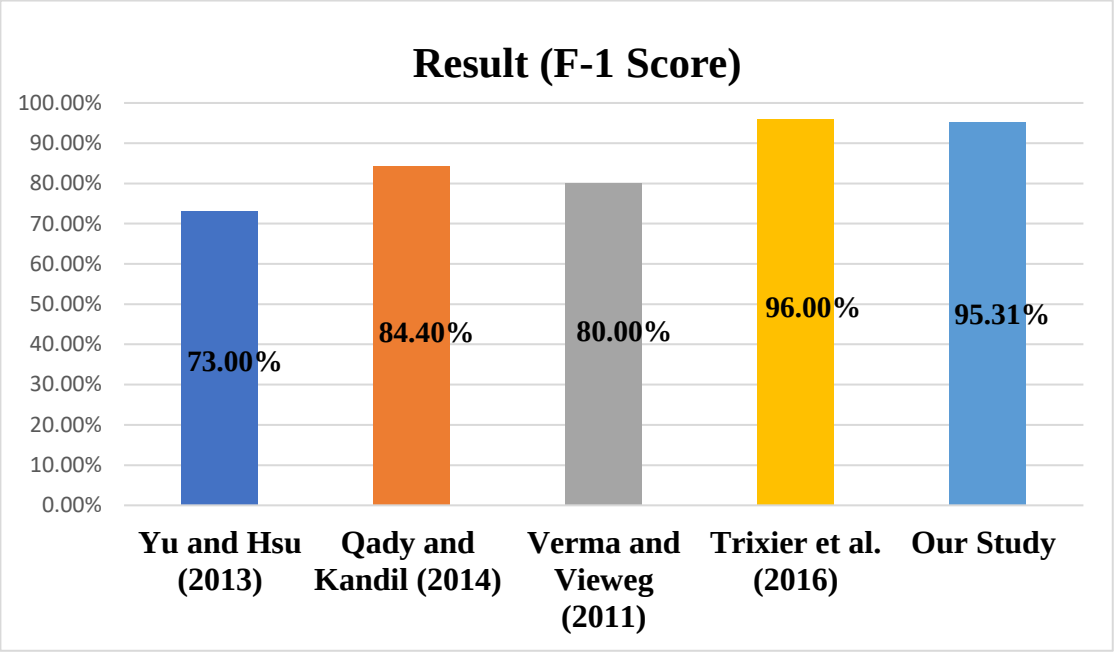

Figure 14 - Comparison of Results from Literature

It may be noted that the manually coded approach's outperforming all the referred ML-based systems is not new as Sagae and Lavie have concluded the same in their study that hand-coded rules enable authors to transmit their intellect and expertise into the coding process (Sagae and Lavie, 2003). For deeper understanding and much better results, manual rule-based systems rather than statistical machine learning tools are preferred especially in the cases where the application domain is rather specific than general (Wang1 *et al.*, 2002). The detailed output outlining the required values for calculation of the F-1 score is given in Figure 15.

Calcualtion of F-1 Score	
Total Outputs	815
True Negative (TN)	266
False Positive (FP)	24.2
Flase Negative (FN)	25
True Positive (TP)	499.8
Precision	95.382%
Recall	95.236%
F-1 Score	95.309%

Figure 15- Calculation of F-1 Score

The NLP system effectively solves the critical constraints of MCA and IE as listen down in Table 6. To start with, it is time-efficient as it can process hundreds of letters within seconds and thus saving on time-related costs as well. It can easily handle the massive amount of input files thus effectively analyzing massive data. In addition, the precision and recall being so high for the results signify that the output is considerably accurate. The table populated by the system is consistent as the letters are assessed and then summarized in the decided headers. By automating the process considerably manual input has been reduced this minimizing impact of personal skill level, increasing work efficiency and, less data duplication. The output sheet populated is reusable and can be used for visual representation. The data is maintained effectively, and precision values signify that loss of data is less.

Table 6 - MCA's Critical Inefficiencies Catered by NLP

S. No.	Critical Constraints of MCA and IE	Benefits of NLP- System
1	Time-Consuming in processing	Time Efficient (60 letters processed in 2 seconds)
2	Costly	Avoid Cost Over-Runs
3	Difficulty in Massive Data Analysis	Capable of handling mounds of Data
4	Complexity of text retrieval	Better Data Analytics and Identification
5	Lack of Structure and Consistency of data	Same Structure as Headers are same
6	Error-Prone in Large data	Increased Accuracy in Comprehension
7	Impact of Personal Skill level	Human Input decreased considerably
8	Reduced Work Efficiency	Work efficiency constant
9	Slow Update of Information	Quick Summarization of Information
10	Reduced Reusability (Scope of application)	Data tabulated in Structured Form
11	Relevancy (Reduced objectivity)	Objective Analysis as human input decreased
12	Data Re-entry	Data Entered by the computer once
13	Inadequate Data Maintenance	Effective Maintenance of Data as Summary in a tabulated form
14	Data loss (Storage Problems)	Data Storage Easier in Electronic Form
15	Data Redundancy/ Duplication of Data	Data Extraction based on headers so lesser duplication
16	Ineffective Visual Representation	Tabular data is easily visually represented

Furthermore, the assessment of the NLP system and in turn the automated framework was done by putting the question in front of five field professionals. They were asked to rate whether the system has successfully avoided a certain inefficiency of MCA and IE on a Likert scale. The score was from 1 to 5, where five described that the system has perfectly catered for the said problem, where in contrast, one shows that the system is incapable of averting the said problem. The survey revealed that except for data redundancy and duplication, where the opinion remained neutral, all other constraints of MCA and IE were believed to be catered for by the developed system. The highest score in 5 was 4.8, whereas the lowest score was 3.2. The

average score came out to be 4.23 signifying that the experts believe the issues of the traditional method that were identified in chapter 3 are successfully avoided by the proposed NLP - based framework and the developed tool.

Summarizing the above discussion, the case study was based on a real project-based sixty letters taken from an ongoing dam project, and its output was formulated by the created NLP – System. This output was given to experts for review and identification of metrics used for the calculation of the F-1 score. The score came out to be above Ninety-Five (95) percent, thus signifying that our proposed Framework can effectively extract information from construction correspondence. The system has numerous advantages over the manual extraction of information, such as time-effective, cost-saving, relevant data extraction, handling large quantities of data, and others.

Chapter 5

CONCLUSION AND RECOMMENDATIONS

This chapter shall summarize the discussion given in the dissertation and list out the limitations while conducting this study. Additionally, it will give recommendations for the better evaluation and formation of the NLP System. It ends with suggestions for future research.

5.1 CONCLUSION

The corporate world as a whole and the construction industry in specific has been using textual data as a means of conveying useful information in day-to-day correspondence, meeting minutes, interim reports, project dashboards, presentations, emails, and others. Mounds and mounds of data are piled up for analysis by the persons concerned and thus leading to a plethora of critical issues such as excessive time consumption, costs incurred, the complexity of retrieval of data, reusability, errors, etc. Automatic content analysis can be used to cater to these inefficiencies.

This research tested the idea that the needs of manual content analysis and information extraction can be considered fulfilled by using NLP based system. The prototype developed was in Python language making use of NLP techniques such as tokenization, parsing, and POS-tagging which was used to summarize 60 letters from a dam project chiefly containing contractual content. After validating the results from Six field experts the precision and recall came out to be 95.38 % and 95.23 % and an F-1 score of above Ninety-five percent i.e., 95.31 %.

5.2 LIMITATIONS

First and foremost, the developed system is limited by employing manually coded complex extraction guidelines. It is not adept or robust enough to cater to input that is erroneous or unfamiliar in itself. Missing, misspelled, and unseen words are cases in point as they directly affect the system's output in terms of extraction of information. In essence, the system’s result is highly influenced by the quality of textual data. Secondly, access to actual project-based letters which were divided into separate categories according to their purpose was not easy. At the end, the

unavailability of data in text or word format was an issue as most of them were in pdf format. For an extensive evaluation, a data scientist is necessary to continuously update the rules. Selection of NLP technique is also a big decision and parsing was adopted as it served the objective of the corpus given but for a detailed application, two or three techniques in tandem with this should be applied, and then a recommendation should be made.

5.3 RECOMMENDATIONS

The Framework should be assessed by applying to a project from inception to completion to get a better idea of its weaknesses for improvement and further refining of rules. It should also be applied and updated for internal communication in bigger organizations where the fixing of responsibility should be assessed based on the content of a letter as evidence from the noting in various departments.

The case study made use of contractual and management letters only whereas correspondence containing design and specification entailing disciplines of civil engineering like geotechnical, structures, water resource, etc. can also be assessed by adding their subject-specific vocabulary and rules.

5.4 FUTURE RESEARCH

A method in which the field of AI optical character recognition (OCR) technology is used to extract any information in pdf or scanned image format and then fed it to an NLP system is much needed. This can completely fulfill the need of the industry.

Moreover, an automated answer or a reply generation software that forms its basis on the data extracted by an NLP system will be an invaluable addition to the construction industry, which is rapidly evolving and exploring new options to introduce information computer technologies.

REFERENCES

Abbaszadegan, A. and Grau, D. (2015) 'Assessing the Influence of Automated Data Analytics on Cost and Schedule Performance', *Procedia Engineering*, 123, pp. 3–6. doi: 10.1016/j.proeng.2015.10.047.

Alashwal, A. M., Rahman, H. A. and Beksin, A. M. (2011) 'Knowledge sharing in a fragmented construction industry : On the hindsight', 6(7), pp. 1530–1536. doi: 10.5897/SRE10.645.

Alsubaey, M., Asadi, A. and Makatsoris, H. (2015) 'A Naïve Bayes approach for EWS detection by text mining of unstructured data: A construction project case', *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*. IEEE, pp. 164–168. doi: 10.1109/IntelliSys.2015.7361140.

Bai, X. (2011) 'Predicting consumer sentiments from online text', *Decision Support Systems*. Elsevier BV, 50(4), pp. 732–742. doi: 10.1016/j.dss.2010.08.024.

Barbella, D. *et al.* (2009) 'Understanding Support Vector Machine Classifications via a Recommender System-Like Approach.', in *DMIN*. Las Vegas, NV, pp. 305–311.

Berg, B. (2004) 'Qualitative Research Methods for the Social Sciences 5', *Teaching Sociology*, 18. doi: 10.2307/1317652.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Brill, E. and Mooney, R. J. (1997) 'An Overview of Empirical Natural Language Processing', 18(4), pp. 13–24.

Caldas, C. H., Soibelman, L. and Han, J. (2002a) 'Automated classification of construction project documents', *Journal of Computing in Civil Engineering*, 16(4), pp. 234–243. doi: 10.1061/(ASCE)0887-3801(2002)16:4(234).

Caldas, C. H., Soibelman, L. and Han, J. (2002b) 'Automated Classification of Construction Project Documents', *Journal of Computing in Civil Engineering*, 16(4), pp. 234–243. doi: 10.1061/(asce)0887-3801(2002)16:4(234).

Cherpas, C. (1992) ‘Natural Language Processing , Pragmatics , and -Verbal Behavior’, pp. 135–147.

Chiticariu, L., Li, Y. and Reiss, F. R. (2013) ‘Rule-based information extraction is dead! Long live rule-based information extraction systems!’, *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (October), pp. 827–832.

Cunningham, H. *et al.* (2011) ‘Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science; 2011’.

Dörre, Jochen; Gerst, Peterl; Seiffert, R. (1999) ‘1999. Text mining finding nuggets in mountains of textual data.pdf’.

Duncan, D. F. (1989) ‘Content analysis in health education research: An introduction to purposes and methods’, *Health Education*, 20(7), pp. 27–31. doi: 10.1080/00970050.1989.10610182.

Eastman, C. *et al.* (2009) ‘Automatic rule-based checking of building designs’, *Automation in Construction*. Elsevier BV, 18(8), pp. 1011–1033. doi: 10.1016/j.autcon.2009.07.002.

ESMAEILI, B. ; H. and Matthew (2012) ‘Attribute-based Risk Model for Measuring Safety Risk of Struck-by Accidents’, *Construction Research Congress*, (Mathiassen 1993), pp. 778–786.

Evans, M., Mcintosh, W., *et al.* (2007) ‘Recounting the Courts ? Applying Automated Content Analysis to Enhance Empirical Legal Research’, 4(4), pp. 1007–1039.

Evans, M., McIntosh, W., *et al.* (2007) ‘Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research’, *Journal of Empirical Legal Studies*, 4(4), pp. 1007–1039. doi: 10.1111/j.1740-1461.2007.00113.x.

Fan, H. and Li, H. (2013) ‘Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques’, *Automation in Construction*. Elsevier BV, 34, pp. 85–91. doi: 10.1016/j.autcon.2012.10.014.

Feldman, R. *et al.* (1998) ‘Knowledge Management: A Text Mining Approach’, *Proc of the 2nd Int Conf on Practical Aspects of Knowledge Management (PAKM98*, Basel, Swi(April 2016), pp. 1–10. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.7128&rep=rep1&type=pdf.

FIDIC (no date) *FIDIC Terms Glossary*. Available at: https://fidic.org/other-resources/fidic-terms-glossary (Accessed: 10 August 2021).

Fonseca, M. J. and Jorge, J. A. (2003) ‘Towards content-based retrieval of technical drawings through high-dimensional indexing’, *Computers and Graphics (Pergamon)*, 27(1), pp. 61–69. doi: 10.1016/S0097-8493(02)00244-3.

Garbharran, H. and Govender, J. (2012) ‘Critical success factors influencing project success in the construction industry’, pp. 90–108.

Gibson, C. B. and Cohen, S. G. (2003) *Virtual Teams That Work: Creating Conditions for Virtual Team Effectiveness*.

Goh, Y. M. and Ubeynarayana, C. U. (2017) ‘Construction accident narrative classification: An evaluation of text mining techniques’, *Accident Analysis and Prevention*, 108(August), pp. 122–130. doi: 10.1016/j.aap.2017.08.026.

De Graaf, R. and Van Der Vossen, R. (2013) ‘Bits versus brains in content analysis. Comparing the advantages and disadvantages of manual and automated methods for content analysis’, *Communications*, 38(4), pp. 433–443. doi: 10.1515/commun-2013-0025.

Graaf, R. De and Vossen, R. Van Der (2013) ‘Bits versus brains in content analysis . Comparing the advantages and disadvantages of manual and automated methods for content analysis 1 Introduction’, 38(4), pp. 433–442. doi: 10.1515/commun-2013-0025.

Grimmer, J. and Stewart, B. M. (2013) ‘Text as data: The promise and pitfalls of automatic content analysis methods for political texts’, *Political Analysis*. Cambridge University Press, 21(3), pp. 267–297. doi: 10.1093/pan/mps028.

Grossman, D. and Frieder, O. (2004) ‘Information retrieval. Algorithms and

heuristics. 2nd ed'.

Guest, G., Bunce, A. and Johnson, L. (2006) 'How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability', *Field Methods*, 18(1), pp. 59–82. doi: 10.1177/1525822X05279903.

Hindle, D. (1989) 'Acquiring Disambiguation Rules from Text', in *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL '89), pp. 118–125. doi: 10.3115/981623.981638.

Hollings, B. C. and Centre, A. (1999) 'Critical Success Factors for Different Project Objectives', 125(June), pp. 142–150.

Jallan, Y. *et al.* (2019) 'Application of Natural Language Processing and Text Mining to Identify Patterns in Construction-Defect Litigation Cases', *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 11(4), p. 04519024. doi: 10.1061/(asce)la.1943-4170.0000308.

Konchar, M., Sanvido, V. and Members, A. (1998) 'COMPARISON OF US PROJECT DELIVERY SYSTEMS', 124(December), pp. 435–444.

Kondracki, N. L., Wellman, N. S. and Amundson, D. R. (2002) 'Content analysis: Review of methods and their applications in nutrition education', *Journal of Nutrition Education and Behavior*, 34(4), pp. 224–230. doi: 10.1016/S1499-4046(06)60097-3.

LAVER, M., BENOIT, K. and GARRY, J. (2003) 'Extracting Policy Positions from Political Texts Using Words as Data', *American Political Science Review*. Cambridge University Press, 97(2), pp. 311–331. doi: 10.1017/S0003055403000698.

Lee, J., Yi, J. S. and Son, J. (2019) 'Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP', *Journal of Computing in Civil Engineering*, 33(3), pp. 1–13. doi: 10.1061/(ASCE)CP.1943-5487.0000807.

Liddy, E. D. (2001) 'Natural Language Processing'.

Lin, Y. C. and Su, Y. C. (2013) ‘Developing mobile- and BIM-based integrated visual facility maintenance management system’, *The Scientific World Journal*, 2013. doi: 10.1155/2013/124249.

Lin, Y. C., Su, Y. C. and Chen, Y. P. (2014) ‘Developing mobile BIM/2D barcode-based automated facility management system’, *Scientific World Journal*, 2014. doi: 10.1155/2014/374735.

Mani, G. F., Feniosky, P. M. and Savarese, S. (2009) ‘D4AR-A 4-dimensional augmented reality model for automating construction progress monitoring data collection, processing and communication’, *Electronic Journal of Information Technology in Construction*, 14(June), pp. 129–153.

Manning, C. D. and Schütze, H. (1999) *Foundations of statistical natural language processing*. MIT press.

Martínez-Rojas, M., Marin, N. and Amparo Vila, M. (2012) ‘The Role of Information Technologies to Address Data Handling in Construction Project Management’, *Journal of Computing in Civil Engineering*, 30(4), pp. 1–11. doi: 10.1061/(ASCE)CP.1943-5487.

Marzouk, M. and Enaba, M. (2019) ‘Automation in Construction Text analytics to analyze and monitor construction project contract and correspondence’, *Automation in Construction*. Elsevier, 98(December 2017), pp. 265–274. doi: 10.1016/j.autcon.2018.11.018.

Matthes, J. and Kohring, M. (2008) ‘The content analysis of media frames: Toward improving reliability and validity’, *Journal of Communication*, 58(2), pp. 258–279. doi: 10.1111/j.1460-2466.2008.00384.x.

Meer, T. G. L. A. Van Der (2016) ‘Automated content analysis and crisis communication research’, 42, pp. 952–961.

Nedeljkovic, D. and Kovašević, M. (2017) ‘Building a Construction Project Key-Phrase Network from Unstructured Text Documents’, *Journal of Computing in Civil Engineering*, 31(6), pp. 1–14. doi: 10.1061/(ASCE)CP.1943-5487.0000708.

Niknam, M. and Karshenas, S. (2015) ‘Integrating distributed sources of information

for construction cost estimating using Semantic Web and Semantic Web Service technologies', *Automation in Construction*, 57, pp. 222–238. doi: 10.1016/j.autcon.2015.04.003.

Niu, J. and Issa, R. R. A. (2014) 'Rule-based NLP Methodology for Semantic Interpretation of Impact Factors for Construction Claim Cases', *ASCE Comuting in Civil and builiding engineering*, pp. 455–462. doi: 10.1061/9780784413616.053.

Al Qady, M. and Kandil, A. (2013) 'Document management in construction: Practices and opinions', *Journal of Construction Engineering and Management*, 139(10), pp. 1–7. doi: 10.1061/(ASCE)CO.1943-7862.0000741.

Al Qady, M. and Kandil, A. (2014a) 'Automatic clustering of construction project documents based on textual similarity', *Automation in Construction*. Elsevier BV, 42, pp. 36–49. doi: 10.1016/j.autcon.2014.02.006.

Al Qady, M. and Kandil, A. (2014b) 'Automatic clustering of construction project documents based on textual similarity', *Automation in Construction*, 42, pp. 36–49. doi: 10.1016/j.autcon.2014.02.006.

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

Sagae, K. and Lavie, A. (2003) 'Combining rule-based and data-driven techniques for grammatical relation extraction in spoken language', in *Proceedings of the Eighth International Conference on Parsing Technologies*.

Sager, N. (1988) 'Medical Language Processing: Computer Management of Narrative Data by Naomi Sager, Carol Friedman, and Margaret S. Lyman (Addison-Wesley 1987)', *SIGCHI Bull.* New York, NY, USA: Association for Computing Machinery, 20(1), pp. 70–71. doi: 10.1145/49103.1046397.

Salminen, J. *et al.* (2019) 'Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type', *Journal of Business Research*, 101(September 2018), pp. 203–217. doi: 10.1016/j.jbusres.2019.04.018.

Soibelman, L. *et al.* (2008) ‘Management and analysis of unstructured construction data types’, *Advanced Engineering Informatics*, 22(1), pp. 15–27. doi: https://doi.org/10.1016/j.aei.2007.08.011.

Stackman, R. W. and Henderson, L. S. (2010) ‘An Exploratory Study of Gender in Project Management : Interrelationships’, 41(5), pp. 37–55. doi: 10.1002/pmj.

Stemler, S. (2001) ‘An overview of content analysis’, *Practical Assessment, Research and Evaluation*, 7(17), pp. 1–10. doi: 10.1362/146934703771910080.

Sullivan, D. (2001) *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*. New York, NY, USA: John Wiley & Sons, Inc.

Tierney, P. J. (2012) ‘A qualitative analysis framework using natural language processing and graph theory’, *International Review of Research in Open and Distance Learning*, 13(5), pp. 173–189. doi: 10.19173/irrodl.v13i5.1240.

Tixier, A. J. *et al.* (2020) ‘Automation in Construction Automated content analysis for construction safety : A natural language processing system to extract precursors and outcomes from unstructured injury reports’, *Automation in Construction*. Elsevier BV, 62(2016), pp. 45–56. doi: 10.1016/j.autcon.2015.11.001.

Tixier, A. J. P. *et al.* (2016) ‘Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports’, *Automation in Construction*. Elsevier BV, 62(2016), pp. 45–56. doi: 10.1016/j.autcon.2015.11.001.

Ur-Rahman, N. (2017) ‘Textual Data Mining For Knowledge Discovery and Data Classification: A Comparative Study’, *European Scientific Journal, ESJ*, 13(21), p. 429. doi: 10.19044/esj.2017.v13n21p429.

Ur-Rahman, N. and Harding, J. A. (2012) ‘Textual data mining for industrial knowledge management and text classification: A business oriented approach’, *Expert Systems with Applications*. Elsevier Ltd, 39(5), pp. 4729–4739. doi: 10.1016/j.eswa.2011.09.124.

Verma, Sudha; Vieweg, S. (2011) ‘Natural Language Processing to the rescue?

Extracting “ Situational Awareness” Tweets During Mass Emergency’, *Circulation*, 2(6), pp. 900–906. doi: 10.1161/01.CIR.2.6.900.

Villanova, M. P. (2014) ‘Attribute-based Risk Model for Assessing Risk to Industrial Construction Tasks’, *ProQuest Dissertations and Theses*, p. 51. Available at: http://ezproxy.library.ubc.ca/login?url=https://search.proquest.com/docview/1655818437?accountid=14656%0Ahttp://gw2jh3xr2c.search.serialssolutions.com/directLink?&atitle=Attribute-based+Risk+Model+for+Assessing+Risk+to+Industrial+Construction+Tasks&author.

Wallach, H. M. *et al.* (2009) ‘Evaluation Methods for Topic Models’, in *Proceedings of the 26th Annual International Conference on Machine Learning*. New York, NY, USA: Association for Computing Machinery (ICML ’09), pp. 1105–1112. doi: 10.1145/1553374.1553515.

Wang, L. C. (2008) ‘Enhancing construction quality inspection and management using RFID technology’, *Automation in Construction*, 17(4), pp. 467–479. doi: 10.1016/j.autcon.2007.08.005.

Wang, X. *et al.* (2018) ‘Improving Workplace Hazard Identification Performance Using Data Mining’, *Journal of Construction Engineering and Management*, 144(8). doi: 10.1061/(ASCE)CO.1943-7862.0001505.

Wang1, Y.-Y. *et al.* (2002) ‘Combination of statistical and rule-based approaches for spoken language understanding’, in *Seventh International Conference on Spoken Language Processing*.

Wiesenfeld, B. M. *et al.* (1999) ‘Communication Patterns as Determinants of Organizational Identification in a Virtual Organization Communication Patterns as Determinants of Organizational Identification in a Virtual Organization’, (July 2019).

Yoon, B. and Park, Y. (2004) ‘A text-mining-based patent network: Analytical tool for high-technology trend’, *Journal of High Technology Management Research*, 15(1), pp. 37–50. doi: 10.1016/j.hitech.2003.09.003.

Yu, W. Der and Hsu, J. Y. (2013a) ‘Content-based text mining technique for retrieval of CAD documents’, *Automation in Construction*, 31, pp. 65–74. doi: 10.1016/j.autcon.2012.11.037.

Yu, W. Der and Hsu, J. Y. (2013b) ‘Content-based text mining technique for retrieval of CAD documents’, *Automation in Construction*. Elsevier BV, 31, pp. 65–74. doi: 10.1016/j.autcon.2012.11.037.

Yu, W. and Hsu, J. (2013) ‘Automation in Construction Content-based text mining technique for retrieval of CAD documents’, 31, pp. 65–74.

Zhang, J. and El-gohary, N. M. (2012) *Extraction of Construction Regulatory Requirements from Textual Documents Using Natural Language Processing Techniques*. doi: 10.1061/9780784412343.0057.

Zhang, J. and El-gohary, N. M. (2016) ‘Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking’, 30(2016), pp. 1–14. doi: 10.1061/(ASCE)CP.1943-5487.0000346.

Zhang, J., El-gohary, N. M. and Asce, A. M. (2016) ‘Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking’, 30(Doe 2011), pp. 1–14. doi: 10.1061/(ASCE)CP.1943-5487.0000346.

Zwikael, O. (2009) ‘Critical planning processes in construction projects’. doi: 10.1108/14714170910995921.