Framework for Automated Information Extraction from Meeting Minutes

Final Year Project UG 2018

By

Hadee Farhan	00000 255815
Maaz Ahmed	00000 253390
Ossam Sikander	00000 242075
Awais Imtiaz	00000 244959

NUST Institute of Civil Engineering
School of Civil and Environmental Engineering
National University of Sciences and Technology, Islamabad, Pakistan

2022

This is to certify that

Final Year Project titled

“Framework for Automated Information Extraction from Meeting Minutes”

Submitted By

Hadee Farhan	00000 255815
Maaz Ahmed	00000 253390
Ossam Sikander	00000 242075
Awais Imtiaz	00000 244959

has been accepted towards the requirements

for the undergraduate degree

in

CIVIL ENGINEERING

Dr. Omer Zubair

Associate Professor
School of Civil and Environmental Engineering
National University of Sciences and Technology, Islamabad

ACKNOWLEDGEMENTS

In the name of Allah, the most Beneficent and the most Merciful as well as due salutations upon Prophet Muhammad (P.B.U.H).

We would like to thank Allah for giving us the opportunity to carry out this research, because nothing happens unless He wills it.

We would also like to thank our parents and teachers, whose years of hard work and commitment enabled us to come this far.

Our supervisor, Dr. Omer Zubair provided us with a lot of guidance throughout the project, and we would truly have been lost if not for his directions

And finally, we would like to thank our friends and colleagues who have supported us so much through all walks of life.

Dedicated

To

Our Supervisor Dr. Omer Zubair &

Our Families

ABSTRACT

The construction industry is highly information dependent, and millions of textual documents are generated everyday within the industry. Meeting meetings are a very crucial part of the entire construction process, with daily, weekly, and monthly progress meetings playing an important role as a key means of correspondence amongst all involved stakeholders. Meeting minutes are generally analyzed and accessed manually, and this process leads to inefficiency in both time and money, as well as being quite prone to errors. Natural Language Processing (NLP) is a linguistic based approach that uses AI and rule-based programing to analyze text. Rule-based NLP was found to be more effective for this project. Over 40 samples of construction progress meeting minutes were obtained, and a standard format was defined. Our program was a Python script, using the samples of construction progress meeting minutes as input and putting it through processes like parsing, tokenization and POS-tagging using manually defined Information extraction rules. The output is in the form of an excel sheet which organizes all the extracted information under appropriate headings. This forms a database where all the analyzes construction progress meeting minutes are stored. This automated process is less prone to errors and delays in time, and is much more cost effective and user-friendly, as well as being more effective as analyzing a large amount of data.

Table of Contents

LIST OF ACRONYMS

IE	Information extraction
NLP	Natural Language Processing
MCA	Manual Content Analysis
AI	Artificial Intelligence
ML	Machine Learning
HSE	Health, Safety and Environment
TM	Text Mining
POS	Parts of Speech

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

We live in an era of information, and information serves as the most vital aspect of the construction industry. It serves to tie together all organizational networks, as well as keep all stakeholders in the loop. This is positively crucial to the success of any project. The traditional way of analyzing documents is manual content analysis. Construction progress meeting minutes, in particular, are a very important form of construction documentation.

1.1 SIGNIFICANCE OF STUDY

The construction industry today is more complex and fragmented than ever, while the processes are becoming more and more technical, and the projects are becoming more complex. In particular, Pakistan's construction industry is really behind on automation, and all information is analyzed manually. About 80 percent of information in the industry is available in textual format. Not being able to effectively and efficiently parse this information can lead to loss of information and more effort. The sheer amount of data available can also lead to information overload, where there is more information than you know what to do with.

All of this can spell disaster for a project. In this dissertation, we take a look at the significance of construction progress meeting minutes, and how the industry analyzes them accesses them manually and how that leads to inefficiency. Then, a solution is proposed using text mining and a framework for automated information extraction to form a database for easy accessing and analysis of construction progress meeting minutes.

1.2 OBJECTIVES

- To identify how manual content analysis of construction progress meeting minutes leads to inefficiency
- To develop an automated framework for information extraction from construction progress meeting minutes

1.3 PROBLEM STATEMENT

Manual Content Analysis and manual Information Extraction is outdated and leads to a lot of inefficiency. This lower efficiency causes people to do extra, unnecessary work and increases the cost of the project. The sheer amount of information generated during a construction project is not conducive to the reusability of that information, and reusability is significantly reduced.

Several research ventures in the past have tried to combat this problem by using advanced computer technologies like Natural Language Processing (NLP) and text mining. This project aims to use these techniques and utilize them to automate the process of extracting meeting minute.

1.4 THESIS STRUCTURE

Introduction is given in the Chapter 1 followed by a thorough literature review in Chapter 2 regarding previous works done on automated content analysis and its uses in real world situations. In Chapter 3, the methodology of the study is discussed. Chapter 4 deals with results of the project, while Chapter 5 discusses the conclusion, limitations and recommendations.

CHAPTER 2

LITERATURE REVIEW

2.1 CONTENT ANALYSIS

Analysis of content happens everywhere in our daily lives. We take in all the information that is read by our senses and analyze it. Analyzing and extracting information from documents is one of the oldest forms of human communication.

Content analysis refers to any technique that analyzes information in a specific manner and retrieves it. It can also be defined as drawing inferences by extracting information from any content that is of interest to the reader. A broad definition of content analysis could be "any technique for making inferences by objectively and systematically identifying specified characteristics of messages." (Stemler, 2001)

Information extraction (IE) is simply using pre-defined classification schemes to reduce the content to an organized or tabulated form. This data can now be analyzed quantitatively or qualitatively

2.2 TRADITIONAL METHODS OF INFORMATION EXTRACTION

Content analysis is generally done traditionally, where data is manually input and extracted manually as well. This method involves painstakingly going over all documents, sorting out and them noting all relevant data.

Meeting minutes are also generally recorded manually and when they need to be accessed, it is a time-taking procedure. Effective information management becomes very difficult with manual content entry and manual content analysis.

Manual content analysis is a very time-inefficient process and manually extracting all that information and tabulating it takes time (Kondracki, Wellman, and Amundson, 2002)

2.3 AUTOMATED INFORMATION EXTRACTION AND CONTENT ANALYSIS

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. The field of technology that deals with automated content analysis is text mining. With the application of specialized techniques, text mining can automate the process of information extraction.

2.3.1 Natural Language Processing (NLP) and Text Mining

Text mining is an automated process where a collection of texts is analyzed using certain analysis tools to extract some sort of information from those texts. Text mining uses techniques such as Natural Language Processing (NLP), Information Retrieval and Data Mining.

There is a need for a tool that helps in the analysis of data from a large amount of structed or unstructured documents, and Natural Language Processing (NLP) fulfills this need as a novel technique that is used to extract relevant data from a large number of documents, in the relevant domain. How it works is that assigns a collection of tags to each part of each document, and then various operations are performed on these tags. Data is parsed and keywords are matched, and various language processing techniques help the program identify the role that different words and sentences are fulfilling. After this identification, the data can be split into different categories and categorized according to the need of the user. Due to this utility of text mining, interest in this field has increased significantly in recent times.

Natural Language Processing (NLP) is a technique that is utilized in text mining. Traditional forms of text mining, though they are effective, still have difficulty in correctly interpreting the nuances of the human language. To overcome this shortcoming, we use techniques such as Natural Language processing (NLP). Natural

Language Processing uses artificial intelligence and linguistics to correctly interpret things in the human language, thereby enabling a computer to understand the human tongue.

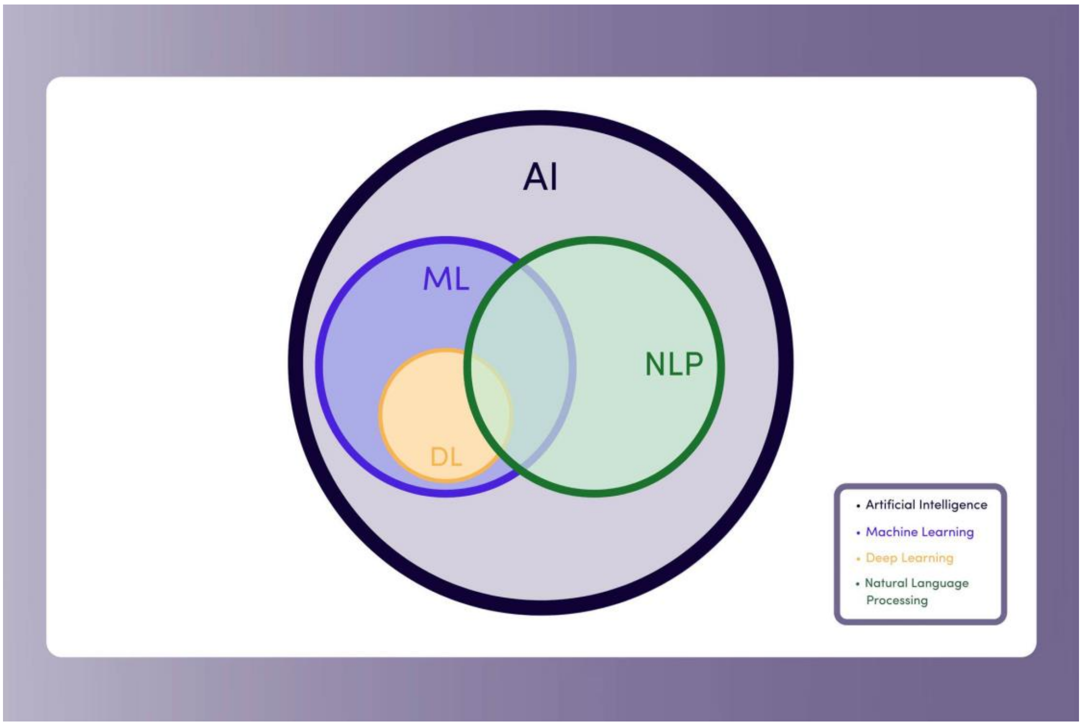


Figure 1 Explanation of NLP

Natural Language Processing relies on ontology, which are niche keywords different for each domain. It is basically a database consisting of the vocabulary of that domain. This causes the program to analyze the text or document based on semantics, which ultimately leads to a much better analysis. Information Extraction (IE) is a sub-field of Natural Language Processing (NLP), where information is extracted from a document according to defined rules and keywords and placed into a template that is better for readability as well as information organization, such as a spreadsheet.

Natural Language Processing is a blessing in an increasingly digitized world where all industries are burdened by a vast amount of digital data, which may or may not be structured. An increasing number of industries are turned towards automated content analysis so that useful information may be extracted and utilized, while keeping labor costs low. Natural Language processing as a concept has many issues though. Human language is very complex and subjective. It is difficult for humans themselves to correctly interpret everything in the human tongue, which is why so many problems stem from miscommunication. Asking a program or a computer to do so is even more difficult. It is very difficult to model rules of grammar as they may vary or may be supremely complex. Likewise, context matters a lot in human language. The same word may have a plethora of meanings when used in different contexts. To solve these problems, Natural Language Processing may utilize techniques such as artificial intelligence and machine learning.

2.3.2 Types of Natural Language Processing

Majorly, two types of approaches are used in the context of Natural Language Processing. One is the rule-based approach and the other is the machine learning based approach.

In the rule-based approach, there are manually coded guidelines which tell the program what keywords to look out for, how to parse data and the how to classify it. These generally work on a trial basis, where to continuously tweak the guidelines until they suit your needs. On the other hand, the machine learning based approach makes use of machine learning algorithms that learn themselves and process the data based on what they have learned. The machine learning based approach needs a very large amount of data. You feed a vast amount of data related to the subject matter, and it trains itself on how to interpret and manage that data, based on such a large sample set.

Due to the large amount of data used in the machine learning based approach, a lot of the times the rule-based approach is used. The use of rule-based Natural Language Processing incorporates an element of human involvement and expertise (Sagae and Lavie, 2003). With the human involvement in rules-based approach, a deeper understanding and better comprehension is possible.

2.3.3 Natural Language Processing and the Construction Industry

Using text mining in the context of construction documents has been done before. The construction industry is one which utilizes vast amounts of data and as such people are always looking for more efficient ways to sort and analyze that data.

Projects in civil engineering are filled with examples where techniques such as data mining and text mining and Natural Language Processing (NLP) were used to perform operations such as the classification and retrieval of data. J. Jeon and X. Xu (2021) from Purdue University took textual specifications for construction quality requirements and used NLP to extricate them. They used a convolutional neural network (CNN), which constitutes a machine learning based approach to Natural Language Processing, utilizing techniques such as sentence classification and syntactic analysis.

P. Jafari and Emad Mohammed (2021) used Natural Language Processing (NLP) and automated extraction to extract reporting requirements from lengthy construction contracts. The study also developed a time-cost prediction framework to identify the time and cost delays associated with preparation of reports.

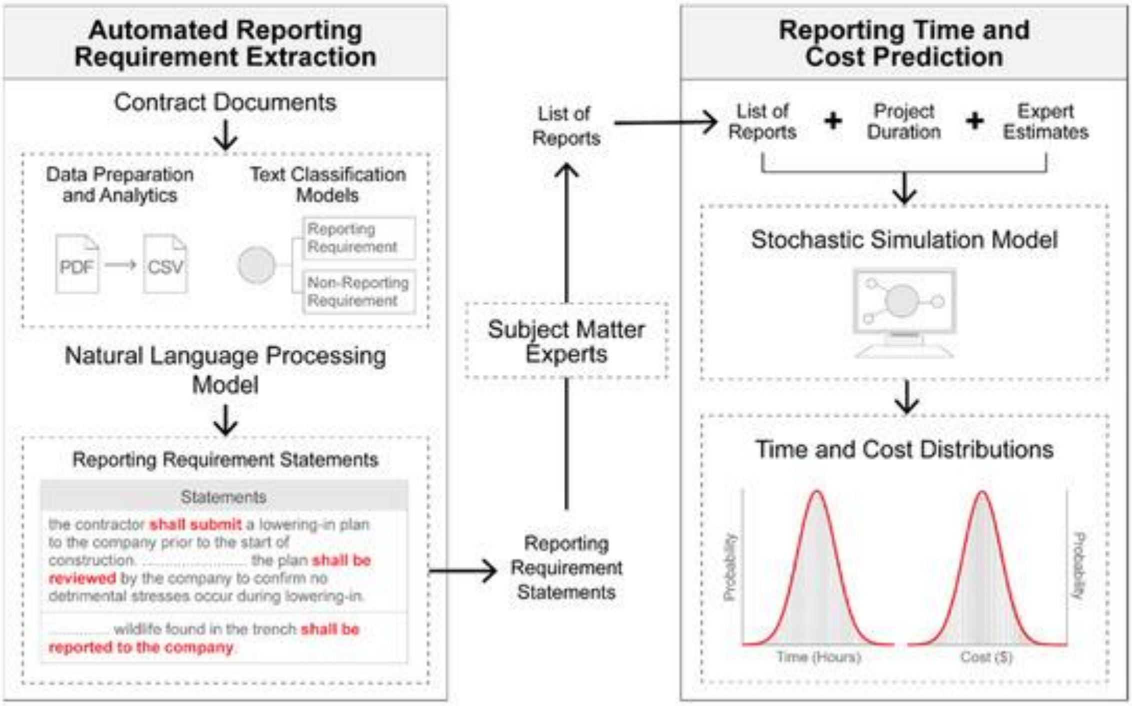

Automated Reporting Requirement Extraction
Contract Documents
Data Preparation and Analytics
Text Classification Models
PDF
CSV
Reporting Requirement
Non-Reporting Requirement
Natural Language Processing Model
Reporting Requirement Statements
Statements
the contractor shall submit a lowering-in plan to the company prior to the start of construction. .... the plan shall be reviewed by the company to confirm no detrimental stresses occur during lowering-in.
.... wildlife found in the trench shall be reported to the company.
List of Reports
Subject Matter Experts
Reporting Requirement Statements
Reporting Time and Cost Prediction
List of Reports
+
Project Duration
+
Expert Estimates
Stochastic Simulation Model
Time and Cost Distributions
Probability
Time (Hours)
Cost ($)
Probability

Figure 2 Layout of study carried out by P. Jafari and Emad Mohammed (2021)

Rule-based Natural Language Processing was used in the automation of checking compliance of regulatory construction documents by Zhang and El-Gohary.

They used IE rules of pattern matching and the manually coded rules were coded using Java software. They also used an F-1 score to validate their results and the efficacy of their tool (Zhang and El-Gohary, 2016)

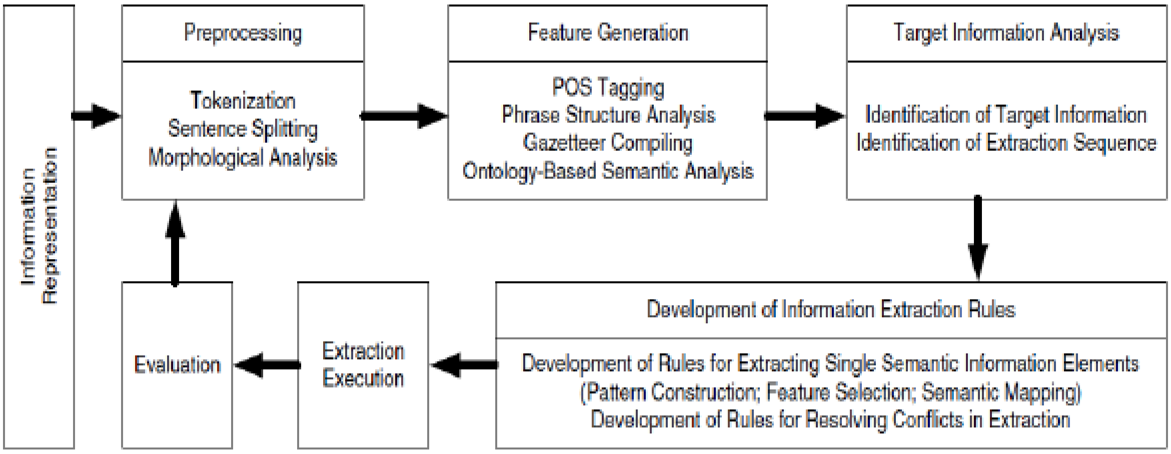

Information Representation
Preprocessing
Tokenization
Sentence Splitting
Morphological Analysis
Feature Generation
POS Tagging
Phrase Structure Analysis
Gazetteer Compiling
Ontology-Based Semantic Analysis
Target Information Analysis
Identification of Target Information
Identification of Extraction Sequence
Development of Information Extraction Rules
Development of Rules for Extracting Single Semantic Information Elements
(Pattern Construction; Feature Selection; Semantic Mapping)
Development of Rules for Resolving Conflicts in Extraction
Extraction Execution
Evaluation

Figure 3 Outline of Rule-based framework used by Zhang and El-Gohary

After going through all these previous studies, it can be concluded that Natural Language processing can, in fact, be a very useful tool in the processing and analyzing of a large amount of human-generated text, relating to the construction industry or otherwise. There is also no significant amount of research done on utilizing Natural Language Processing to extract information from monthly construction progress meeting minutes. This is a domain where further work can be done, and hence this team has decided to take up the task of creating a framework for automated information extraction from construction progress meeting minutes using Natural Language Processing.

CHAPTER 03

METHODOLOGY

This chapter will underline the methodology that was used for this project. This includes the identification of the problem, selection of a solution and then the steps undertaken in order to build our framework and to continuously improve it.

3.1 IDENTIFICATION OF PROBLEM

After a detailed literature review, the issues that accompany manual content analysis were identified which included the overlooking of information, difficulty in accessing and sorting through such a huge number of textual documents, and costs in terms of labor, time and money as well as the inefficiency accompanying manual content analysis.

This problem that this team identified was further validated through consultation with many industry sources which agreed that the construction industry is reliant on text-based correspondence and manually sorting through all this data is indeed time-consuming and inefficient. Automated content analysis was proposed as the solution to this problem.

3.1.1 Research Design

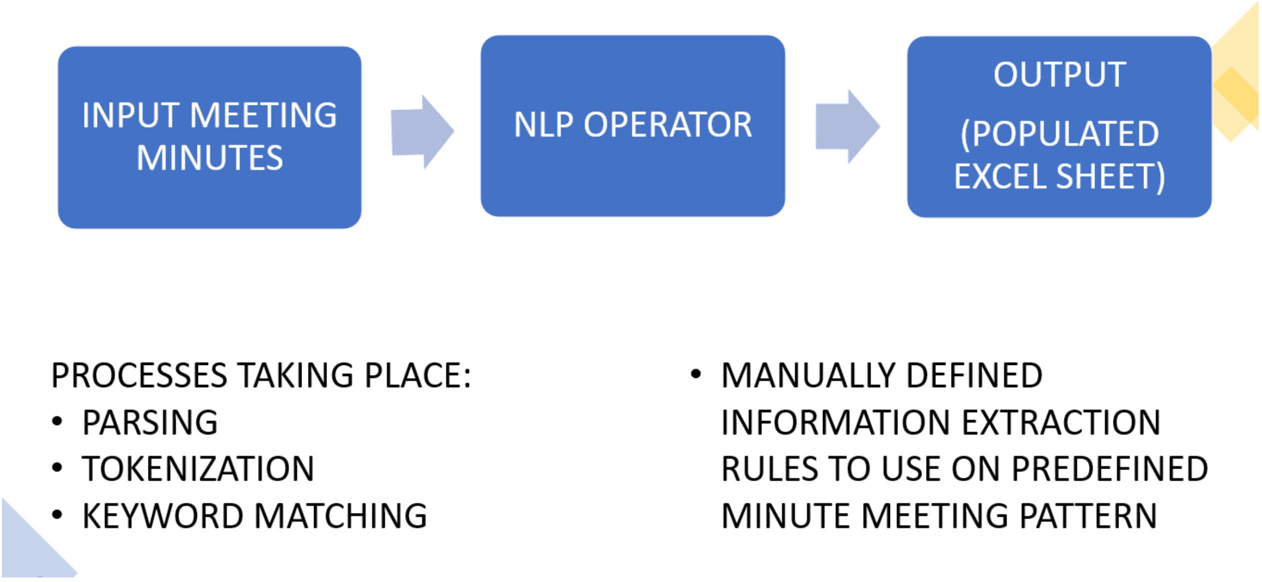
The study and research conducted in the initial stages of the project included discovering the gaps among the studies done previously, helped a great deal in formulating the problem statement and selection of our research objectives. A rule-based Natural Language Processing framework was used in order to deal with the inefficiencies in MCA.

3.1.2 Minutes of Meetings

Specifically, minutes of construction progress meetings were chosen to be the dataset that would accompany this research sue to their importance within the construction industry. These meetings serve as an important tool for coordination among all the stakeholders involved in the project, as well as being a vital resource in communication and the continuous updates that are made to the project. Any and all changes in scope or any other prevailing issue that may hinder the project is discussed and the issue is then dealt with in a manner that all stakeholders are satisfied. The minutes of these meetings serve as important records in cases and may need to be called upon if an of the aforementioned information is needed.

3.2 FORMING A FRAMWORK USING NLP

A rule-based NLP framework was decided for the project because it is more effective in specific use-cases. Through thorough research, all the elements needed for the formulation of the framework were identified. We of course needed our inputs (the construction documents) which for this case we chose to be construction progress meeting minutes. A standard pattern was defined for the meeting minutes because that would cause the framework to be much more efficient. Domain-specific ontology was formulated, where keywords were selected from certain domains to let the program know how to interpret each part of the text. The output was in the form of a spreadsheet file. Below if a small visual summary of how our framework goes about completing its task.



PROCESSES TAKING PLACE:

- PARSING
- TOKENIZATION
- KEYWORD MATCHING
- MANUALLY DEFINED INFORMATION EXTRACTION RULES TO USE ON PREDEFINED MINUTE MEETING PATTERN

Figure 4 Natural Language Processing Framework for Automated Content Analysis

For the formulation of the framework, the Integrated Development Environment (IDE) of Visual Studio Code was used, and for our script we chose Python due to its ease of access and user-friendliness. All terminologies which would help develop a better understanding of the Natural Language Processing framework are given below.

3.2.1 Progress initialization

A compiler or interpreter is the first and foremost important thing that you need when the program is started while, the core components may include libraries required files and variables or any other data sort.

3.2.2 Libraries

Our framework utilizes a number of libraries. Most of the times when talking in programming terminology, a library is meant as a permanent resource that can be utilized by computer programmes in order to conduct the development, execution, and the initialization of a software. Libraries may also contain already compiled codes and functions. Among these codes and functions, certain types of specifications for example for classes and values are also developed. So, the library is a previously written set of commands which we can be called upon whenever needed, and the effort to write those codes again and again is saved. The basic function of a purposes library can also be to eliminate the need of coding again and again for simple tasks, while establishing a database of codes, from which commands can be copied and pasted. The pre-compiled commands can be for simple tasks as well as some complex ones. A Library is unique, which means that different libraries can have different functionalities while having certain restrictions on its usage. On its own, a library might not be anything, but when a program is compiled, the first thing to be integrated is a library. A list of the libraries included in our code are:

- docx2txt
- os
- re
- glob
- parse

- csv
- argparse
- os.path
- openpyxl
- xlsxwriter
- sys
- shutil

3.2.3 Variables

The data that is stored in a programme can be described in two ways. It can be either a user input or it can be locally stored Variable within the program. The second form of data just described is called a variable and it gets stored within the programme. Like libraries, a variable is also unique and also has an identifier, which itself is unique. Variables can also be classified on the basis of their identifiers as well as their datatypes. The value assigned to a variable can be changed easily when the program is executed.

In order to fully understand the working of a variable, think of it as a bag classified to store items of same kind. You can assign more than one value onto a variable and it will be stored in it, until the variable is recalled in the program. The recalling of a variable is only possible if you remember it's unique identifier. This data could be either known or unknown, solely based upon the assignment of value to the variable. Based on the locality of the variable, we classify variables into two categories. One, a global variable which has global scope and is accessible as well as usable all over the program. Their extent is the whole of program runtime and these are mostly not value changing variables. On the other hand, local variables have local scope and their values are changeable.

3.2.4 Data

We can define data as a group or cluster of individual statistical data and discrete facts. Keeping the datum a single variable bearing whatever value, data can be defined as a set of either quantitative or qualitative variable values about a number of objects or persons.

We have only talked about operation of the main program until now, to move further, we can talk about data preparation and data integration.

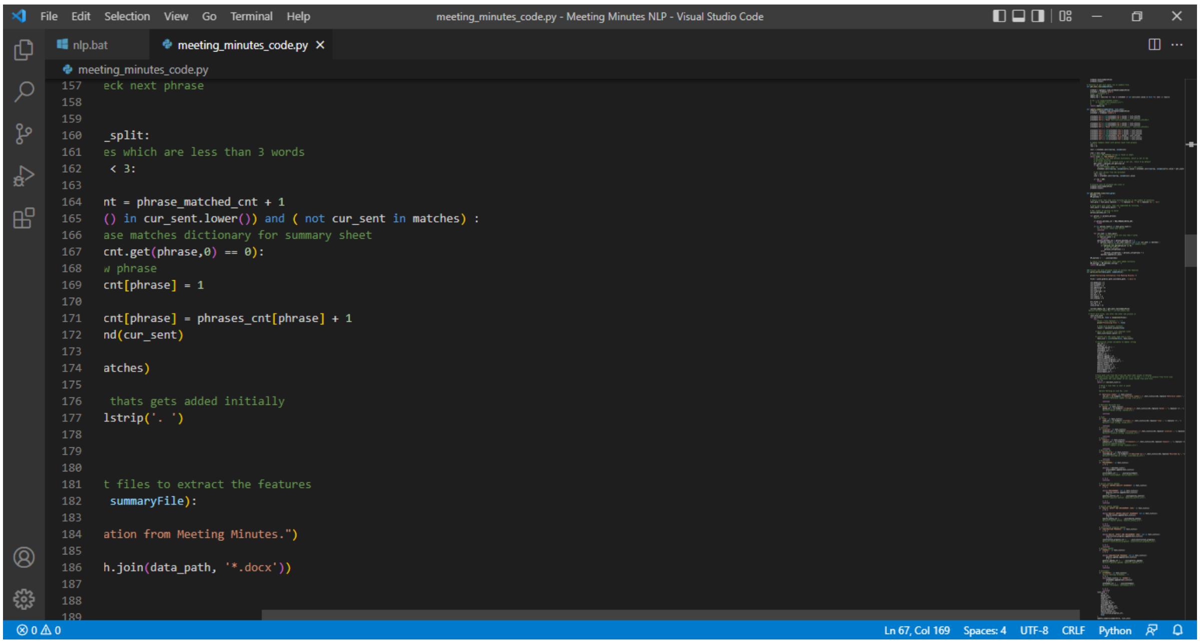
3.2.5 Data availability

A clump or cluster of distinct facts and statistics—often numerical—is referred to as data. A set of values of quantitative or qualitative variables regarding numerous people or items is referred to as data, whereas a datum is a single variable with a single value.

Until now, the preparations for running the main programme were being integrated and implemented. Data will now be included in the process. As soon as they're finished, the main programme gets initialised.

3.3 RUNNING OF THE FRAMEWORK

Now that we have gone over some of the terms that are important to our framework, we turn towards the actual working of the framework itself.



```
eck next phrase


_split:
es which are less than 3 words
 < 3:

nt = phrase_matched_cnt + 1
() in cur_sent.lower()) and ( not cur_sent in matches) :
ase matches dictionary for summary sheet
cnt.get(phrase,0) == 0):
# phrase
cnt[phrase] = 1

cnt[phrase] = phrases_cnt[phrase] + 1
nd(cur_sent)

atches)

 thats gets added initially
lstrip('. ')



t files to extract the features
 summaryFile):

ation from Meeting Minutes.")

h.join(data_path, '*.docx'))


```

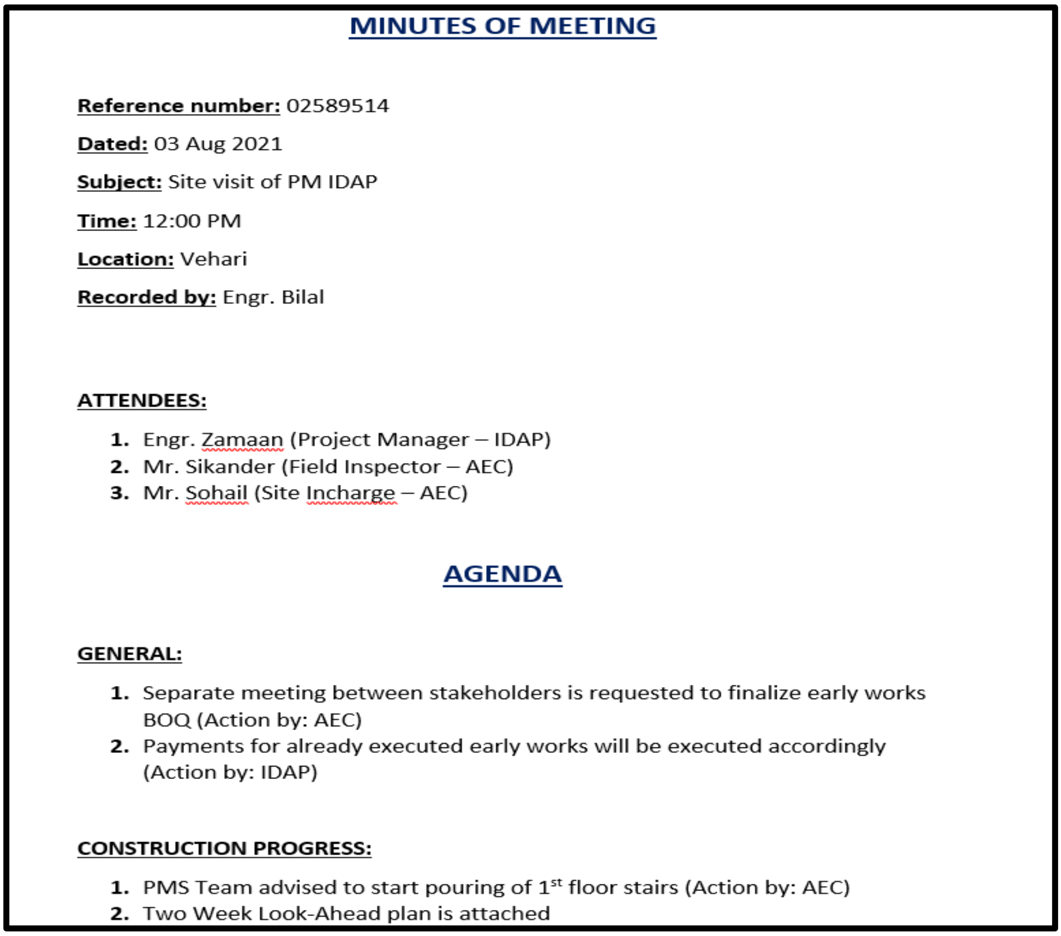
Figure 5 Picture of our Python Script in Visual Studio Code

3.3.1 Data Preparation

The main program can only be initialized after the data preparation has been conducted successfully. We need to make several checks before even starting the main program. These checks may include input directory of files or other files necessary for the operation or running of the program.

3.3.2 Minutes of Meeting

These are the input data sets for our programme to work on and provide a summary of. These are saved in a directory, and the complete path is supplied to the application as an argument. These will be in a directory called “Data” and after processing, they will be transferred to a directory called “Data - Processed” indicating that these letters/files have been processed and the information that might be retrieved has been entered in the summary sheet/file.



MINUTES OF MEETING

Reference number: 02589514

Dated: 03 Aug 2021

Subject: Site visit of PM IDAP

Time: 12:00 PM

Location: Vehari

Recorded by: Engr. Bilal

ATTENDEES:

1. Engr. Zamaan (Project Manager – IDAP)
2. Mr. Sikander (Field Inspector – AEC)
3. Mr. Sohail (Site Incharge – AEC)

AGENDA

GENERAL:

1. Separate meeting between stakeholders is requested to finalize early works BOQ (Action by: AEC)
2. Payments for already executed early works will be executed accordingly (Action by: IDAP)

CONSTRUCTION PROGRESS:

1. PMS Team advised to start pouring of 1st floor stairs (Action by: AEC)
2. Two Week Look-Ahead plan is attached

Figure 6 An Extract from the Standard Meeting Minute Pattern

We had a total of 40 samples of actual construction progress meeting minutes, acquired from various organizations. This acquisition was not easy as meeting minutes are mostly internal documents within organizations which are not made public. This is the reason a larger sample set could not be collected. As mentioned before, a standard pattern was defined for the construction progress meeting minutes, because all organizations use different (often proprietary) formats for recording minutes of meetings. It would have been a major undertaking to enable a program to properly sort through and analyse all these different formats, so a standard format was defined and all the meeting minutes were converted into that standard format.

3.3.3 Analysing Process of the Algorithm

Once the libraries and variables have been initialized, the program begins to process the file content as well as saves it. The program uses the docx2txt library and converts all the textual document into a format that is readable by the program. After this, each line is individually parsed and the program uses the keyword .csv files to search for specific keywords within the text. If a keyword match is positive, the program then tokenizes that line into individual words and performs Parts of Speech-tagging operation, which basically determines what kind of function those words are performing in the context of the text. After all this is done, the program can determine how to classify that piece of text, and it places that text under a certain appropriate heading in the output spreadsheet file. The program performs this operation for all the lines and all the documents in queue. By differentiation of the contents and separating a new content line after splitting lines, all the lines are converted into the list. After this process, the column variables are initialized that helps us in storage of data. While loop is used and data is stored into the column of summary from arrays. Each line is individually processed, as the loop works iteratively until it reaches the last line.

3.3.4 Extraction of Details

The first thing to be extracted from the array is ‘Reference no.’ as they usually come in top portion. It retrieves ‘Reference no.’, ‘Dated’, ‘Subject’, ‘Time’, and ‘Location’. But before saving it in string form to be written on summary file the

punctuation like commas, and tabs are replaced with blank/empty value in the string. Reference extraction is done by scanning the document. At first the punctuation like tabs, multiple spaces, directory value, and other non-useable values are left out. The references are extracted with the help of loop. After getting references it then proceeds to next step, which is to extract Enclosure, if any. Using loop, it goes through all lines to find what else is being sent "encl" or it reaches the end.
'Recorded By', if any, is also done in the same manner of scanning and finding the keywords. Using loop, it goes through all lines to find who else is being sent "copy to" or it reaches the end. If the project name has already not been fetched, through the code in designations loop, there is a loop just to find project when designation is known but not the project. After that program extract the subject of the letter and saves in in a variable.
Now the 'Attendees' details would to be extracted from the data/letter, including name, designation, and organization.

3.3.5 Information Updating

A line is made with all of the columns to be written into results .csv file for data entry. This process occurs only after the information extraction has been carried out.
Values Extracted:

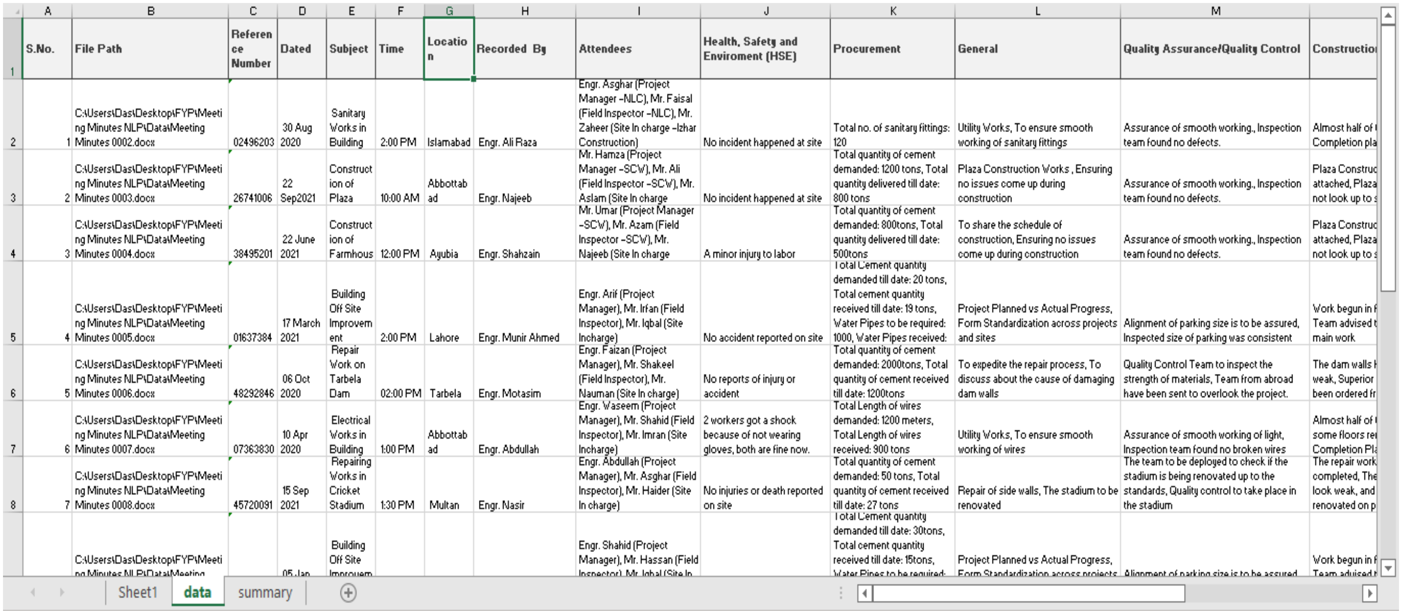
- File Path
- Reference number
- Dated
- Subject
- Time
- Location
- Recorded By
- Attendees
- General
- Construction Progress
- HSE
- Quality Assurance/Quality Control
- Procurement

To summarize the entire process, a code was developed using python script. All input files are stored within an input “Data” folder, and when the program starts, it takes all the files within the input folder and processes them one by one. Natural Language Processing operations are performed on the entire text and all the documents in queue, and once the program is able to understand what function each piece of text is performing, it is able to categorize the text and decide what domain it falls under. It then enters that piece of text under the appropriate heading in the resultant spreadsheet “Extracted Meeting Minutes”. All the files that were in the “Data” folder are, after all processing is done, shifted to the “Data – Processed” folder. In this way, automatic content analysis has been performed on all the meeting minutes, and a comprehensive spreadsheet database has been prepared where all the information is organized.

CHAPTER 04

RESULTS

The results of this study are in the form of an excel sheet. The program that this team developed takes the construction progress meeting minutes in the pre-defined standard format, and after application of rule-based Natural Language Processing the results are displayed in an excel sheet as shown.



S.No.	File Path	Reference Number	Dated	Subject	Time	Location	Recorded By	Attendees	Health, Safety and Enviroment (HSE)	Procurement	General	Quality Assurance/Quality Control	Constructio
1	C:\Users\Das\Desktop\FYP\Meeting Minutes NLP\Data\Meeting Minutes 0002.docx	02496203	30 Aug 2020	Sanitary Works in Building	2:00 PM	Islamabad	Engr. Ali Raza	Engr. Asghar (Project Manager –NLC), Mr. Faisal (Field Inspector –NLC), Mr. Zaheer (Site In charge –Izhar Construction)	No incident happened at site	Total no. of sanitary fittings: 120	Utility Works, To ensure smooth working of sanitary fittings	Assurance of smooth working., Inspection team found no defects.	Almost half of Completion pla
2	C:\Users\Das\Desktop\FYP\Meeting Minutes NLP\Data\Meeting Minutes 0003.docx	26741006	22 Sep2021	Construction of Plaza	10:00 AM	Abbottabad	Engr. Najeeb	Mr. Hamza (Project Manager –SCW), Mr. Ali (Field Inspector –SCW), Mr. Aslam (Site In charge	No incident happened at site	Total quantity of cement demanded: 1200 tons, Total quantity delivered till date: 800 tons	Plaza Construction Works , Ensuring no issues come up during construction	Assurance of smooth working., Inspection team found no defects.	Plaza Construc attached, Plaza not look up to s
3	C:\Users\Das\Desktop\FYP\Meeting Minutes NLP\Data\Meeting Minutes 0004.docx	38495201	22 June 2021	Construction of Farmhous	12:00 PM	Ayubia	Engr. Shahzain	Mr. Umar (Project Manager –SCW), Mr. Azam (Field Inspector –SCW), Mr. Najeeb (Site In charge	A minor injury to labor	Total quantity of cement demanded: 800tons, Total quantity delivered till date: 500tons	To share the schedule of construction, Ensuring no issues come up during construction	Assurance of smooth working., Inspection team found no defects.	Plaza Construc attached, Plaza not look up to s
4	C:\Users\Das\Desktop\FYP\Meeting Minutes NLP\Data\Meeting Minutes 0005.docx	01637384	17 March 2021	Building Off Site Improvement	2:00 PM	Lahore	Engr. Munir Ahmed	Engr. Arif (Project Manager), Mr. Irfan (Field Inspector), Mr. Iqbal (Site Incharge)	No accident reported on site	Total Cement quantity demanded till date: 20 tons, Total cement quantity received till date: 19 tons, Water Pipes to be required: 1000, Water Pipes received:	Project Planned vs Actual Progress, Form Standardization across projects and sites	Alignment of parking size is to be assured, Inspected size of parking was consistent	Work begun in f Team advised t main work
5	C:\Users\Das\Desktop\FYP\Meeting Minutes NLP\Data\Meeting Minutes 0006.docx	48292846	06 Oct 2020	Repair Work on Tarbela Dam	02:00 PM	Tarbela	Engr. Motasim	Engr. Faizan (Project Manager), Mr. Shakeel (Field Inspector), Mr. Nauman (Site In charge)	No reports of injury or accident	Total quantity of cement demanded: 2000tons, Total quantity of cement received till date: 1200tons	To expedite the repair process, To discuss about the cause of damaging dam walls	Quality Control Team to inspect the strength of materials, Team from abroad have been sent to overlook the project.	The dam walls h weak, Superior been ordered fr
6	C:\Users\Das\Desktop\FYP\Meeting Minutes NLP\Data\Meeting Minutes 0007.docx	07363830	10 Apr 2020	Electrical Works in Building	1:00 PM	Abbottabad	Engr. Abdullah	Engr. Waseem (Project Manager), Mr. Shahid (Field Inspector), Mr. Imran (Site Incharge)	2 workers got a shock because of not wearing gloves, both are fine now.	Total Length of wires demanded: 1200 meters, Total Length of wires received: 900 tons	Utility Works, To ensure smooth working of wires	Assurance of smooth working of light, Inspection team found no broken wires	Almost half of some floors re Completion Pla
7	C:\Users\Das\Desktop\FYP\Meeting Minutes NLP\Data\Meeting Minutes 0008.docx	45720091	15 Sep 2021	Repairing Works in Cricket Stadium	1:30 PM	Multan	Engr. Nasir	Engr. Abdullah (Project Manager), Mr. Asghar (Field Inspector), Mr. Haider (Site In charge)	No injuries or death reported on site	Total quantity of cement demanded: 50 tons, Total quantity of cement received till date: 27 tons	Repair of side walls, The stadium to be renovated	The team to be deployed to check if the stadium is being renovated up to the standards, Quality control to take place in the stadium	The repair work completed, The look weak, and renovated on p
	C:\Users\Das\Desktop\FYP\Meeting Minutes NLP\Data\Meeting		05 Jan	Building Off Site Improvem				Engr. Shahid (Project Manager), Mr. Hassan (Field Inspector), Mr. Iqbal (Site In		Total Cement quantity demanded till date: 30tons, Total cement quantity received till date: 15tons, Water Pipes to be required:	Project Planned vs Actual Progress, Form Standardization across projects	Alignment of parking size is to be assured	Work begun in f Team advised t

Figure 7 Output of NLP Framework showing information extracted from meeting minutes

The output file takes the plethora of information in the meeting minutes, and extracts that information and categorizes it. The results show the file path of the document, as well as other crucial information as the reference number of the document, the date, time, subject, list of attendees, the name of the person who recorded the meeting minutes as well as the actual subject matter of the meeting minutes, displayed under various headings such as General, Health Safety and Environment, Construction Progress etc. These headings may seem limiting but are necessary for the implementation of rule-based Natural Language Processing where all guidelines need to be manually fed into the program. The scope of the information that is extracted can be widened very easily, merely by adding a keyword file for the program to call upon and the expanding the number of headings that are recorded.

CHAPTER 5

CONCLUSION

Every enterprise in the modern world has a hoard of textual data that is used for communication or record-keeping, and sorting through or analyzing this data can truly prove a monstrous task. To try and analyze this information which may be in the form of meeting minutes, letters, presentations, reports or miscellaneous records is too tedious to be done manually by humans, and doing so would only lead to information being missed out or overlooked which will result in costs of both time and money. So automatic content analysis could be the answer to this problem.

This study delved into the problems caused by manual content analysis, and proposed a solution using automatic content analysis, specifically for construction progress meeting minutes. A framework was made using the Integrated Development Environment of Visual Studio and the program was coded in the Python script. Using rule-based Natural Language Processing, guidelines and keyword files were given to the program which helped it to parse text, split sentences and determine the purpose and contextual meaning behind each sentence. After this was identified, the program determined what classification to give to the particular piece of text and hence the text was categorically placed under the relevant heading in the output spreadsheet.

The use-case of this study is that this program enables an organization to compile a spreadsheet database of ordered and relevant information contained within construction progress meeting minutes. These minutes are typically analyzed manually and individually, so this database could provide the organization with a central database with ordered information from these meeting minutes which can be easily accessed.

5.1 LIMITATIONS

This study comes with its own set of limitations. Although rule-based Natural Language proves to be more effective in specified use-cases, it still consists of tediously manually coded guidelines. The larger and more complex the input becomes, the more difficult it is to manually code each and every aspect of it. And if you fail to cater to every aspect, the program will have difficulty trying to construe input that is unfamiliar to it. So, the entire system is highly reliant on the quality of the manually coded guidelines as well as the input provided to it. It also requires the rules to continuously be updated. For larger applications, various other automatic content analysis techniques would have to be applied.

Another problem that this team faced was the meeting minutes themselves. Due to their nature, meeting minutes are confidential documents because they may reveal the inner working of an organization as well as information about ongoing projects. As such, organizations typically do not make meeting minutes public and it was not easy to obtain samples for this study. Furthermore, since they are internal documents, every organization may have its own format for meeting minutes. Hence, it was necessary to make a standard format for monthly construction progress meeting minutes and convert all our samples into that standard format. Any organization willing to use this system would have to adopt that standard format for the recording of their meeting minutes.

5.2 RECOMMENDATIONS

If this framework was assessed for the monthly construction progress meeting minutes for an entire construction project from start to finish, a better assessment of its accuracy and efficacy could be made.

Minutes of meetings are just one aspect of the textual records kept in the construction industry. This framework can be scaled and expanded for a variety of construction documentation and even documentation pertaining to other industries, if the relevant keyword files are added, with a slight tweaking of the guidelines.

Another vital use-case that can be achieved by further research inot this study is the use of Artificial Intelligence to completely understand how the records of meeting minutes are kept, and once the format and content is understood by the AI through various Machine Learning algorithms, the AI will be fully capable of recording meeting minutes by itself, hence automating a tedious process and removing the need for excessive manual labor.

REFERENCES

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

2. 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.

3. 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*.

4. Jeon, J., Xu, X., Zhang, Y., Yang, L., & Cai, H. (2021). Extraction of Construction Quality Requirements from Textual Specifications via Natural Language Processing. Transportation Research Record: Journal of the Transportation Research Board, 2675(9), 222–237. https://doi.org/10.1177/03611981211001385

5. Jafari, P.; Al Hattab, M.; Mohamed, E.; AbouRizk, S. Automated Extraction and Time-Cost Prediction of Contractual Reporting Requirements in Construction Using Natural Language Processing and Simulation. *Appl. Sci.* 2021, *11*, 6188. https://doi.org/10.3390/app11136188

6. 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.