FRAMEWORK FOR AUTOMATED INFORMATION EXTRACTION FROM CONSTRUCTION CORRESPONDENCE LETTERS

By:

Muhammad Hassan Mughal (Group Leader)

(NUSTBECE2018&00000258224)

Syed Faaiq Hussain
(NUSTBECE2018&00000267748)
Mazhar Ali Zahid
(NUSTBECE2018&00000241794)

Bachelor of Engineering

In

Civil Engineering

Department Of Construction Engineering and Management

NUST Institute of Civil Engineering (NICE)

School Of Civil and Environmental Engineering (SCEE)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

(2022)

FRAMEWORK FOR AUTOMATED INFORMATION EXTRACTION FROM CONSTRUCTION CORRESPONDENCE LETTERS

By:

Muhammad Hassan Mughal (GL)	NUSTBECE2018&00000258224
Syed Faaiq Hussain	NUSTBECE2018&00000267748
Mazhar Ali Zahid	NUSTBECE2018&00000241794

A thesis submitted to the National University of Sciences and Technology, Islamabad, in partial fulfillment of the requirements for the

Bachelor of Engineering in

Civil Engineering

Thesis Supervisor: Dr. Muhammad Usman Hassan

Co-Supervisor: Dr. Omer Zubair

Department Of Construction Engineering and Management

NUST Institute of Civil Engineering (NICE)

School Of Civil and Environmental Engineering (SCEE)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of BS Thesis written by Mr. Muhammad Hassan Mughal (Registration No. 00000258224), of NUST Institute of Civil Engineering (NICE) has been vetted by undersigned, found complete in all respects as per NUST Statutes/ Regulations/ BS Policy, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of BS degree.

Signature: _____

Name of Supervisor: _____

Date: _____

DECLARATION

It is hereby reverently and truthfully declared that all the work alluded to this thesis is composed by us. Any references to the work done by any other person or University have been appropriately cited.

DEDICATIONS

We would like to dedicate our works to our parents, our teachers our institution NUST, and all our friends. We performed our work with the impressive assurance & resolve, and applied best of ourselves to the errand at hand.

ACKNOWLEDGEMENTS

In the name of Allah, the most Beneficent, the most Merciful; as well as peace and blessings, be upon Prophet Muhammad (S. A. W. W.), Allah's servant and final messenger. We are thankful to Allah almighty for bestowing us an opportunity to be here in a prestigious institute and intellectual strength with continuous guidance to work up to the mark.

We are grateful to our families for unconditional, unequivocal, and loving support.

This work would not have been possible without support of generous faculty of NUST Institute of Civil Engineering & Department of Construction Engineering & Management. We are especially indebted to Dr. Muhammad Usman Hassan, assistant professor at NIT, NUST who provided us with his unending guidance and motivation as supervisor of the project. We are extremely grateful to the university students and faculty members who have helped us directly or indirectly. We are thankful to our department and school for providing us with a leaning environment which helped us in achieving we dreamt of.

TABLE OF CONTENTS

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ACA	Automated Content Analysis
CSV	Comma Separated Values
CA	Content Analysis
DM	Data Mining
DL	Deep Learning
IE	Information Extraction
IR	Information Retrieval
FIDIC	International Federation of Consulting Engineers
KDD	Knowledge Discovery in Databases
ML	Machine Learning
MCA	Manual Content Analysis
NB	Naïve Baes
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OCR	Optical Character Recognition
POS	Parts of Speech
PSG	Phrase Structure Grammar
TM	Text Mining
VSM	Vector Space Model
WIP	Work In Progress

TABLE OF FIGURES

ABSTRACT

Information transfer in construction industry has always been immense mainly in the form of letters, emails or field study notes etc. But the process used predominantly, which is Manual Content Analysis (MCA) or manual method to deal with information extraction has never been quite pleasant due to number of inefficiencies including time-consuming, error prone, costly etc. To refine the approach and enhance the system, automated information extraction was explored and for this a code was prepared consisting of particular libraries and operations. This procedure based on the Natural Language Processing (NLP) technique using Artificial Intelligence (AI) assisted in extracting the useful information out from massive data, in our case the letters.

The positive productivity by using this Automated Content Analysis (ACA) can be observed as it has eliminated all the issues that are associated with or may arise as a result of MCA process by making the Content Analysis (CA) process fast, and less troublesome.

Chapter 01

INTRODUCTION

1.1 BACKGROUND

In a project lifecycle, information or document shifting within department to another is extremely common. For that reason, proper communication handling is a key factor in every corporation. So, if there is any deficiency in this process, loss of useful data might occur and human resource might affect. Advanced methods are required which can replace traditional processes.

1.2 PROBLEM STATEMENT

With manual content analysis, entire documentation is inspected by an individual which end with several flaws in the extracted data. The quality of service is determined by individuals, which could be affected by their frame of mind and energy. There is a strong possibility of passing over any important information without drawing it out, which might cost user losing something expensive.

As multiple information is being entered into the system throughout the day so data swap or one's documents to be misunderstood by another's file is possible. These incidents have a high possibility of occurring in field and can result in losing reputation and networking. Controlling it would consume valuable time and efforts. To eliminate these complications, a framework is to be constructed that would

smoothen up the process and can complete the process in short time. For this an advanced method known as automated content analysis can save a lot of money. Automation has previously been induced in several fields, and has demonstrated convenience and satisfaction to users. Construction industry is also lacking straightforward methods like automation that can make it effortless.

1.3 RESEARCH GAP

All the aforementioned studies that have been previously conducted on this topic have only focused more towards specific areas; mainly the extra works. which is considerable contribution to info extraction in construction industry. But they don't focus on the other aspects like the delays, budgetary changes, what type, where from, where is its addressed to, in our research we'll try to close this gap by focusing and while not forgetting one of the main aspects which is mentioned above as extra works. And the other more sophisticated studies that have been done in the past are not related to the construction industry.

1.4 RESEARCH OBJECTIVES

- To determine inefficiencies in information extraction from construction correspondence letters using traditional content analysis techniques.
- To create a framework for extracting information from construction correspondence letters in a refined and efficient manner.
- To assess the effectiveness of the Framework that has been developed.

1.5 SIGNIFICANCE OF STUDY

In today's world, the construction business is characterized by segmentation, breakdown, and continuing complication in activities and processes. This can limit a project's growth and success by threatening the achievement of important project objectives. To help alleviate this serious situation, increasing interaction and cooperation among diverse industry units can boost productivity and efficiency. Accurate and current information access is crucial in today’s era for the development of cooperative mechanisms and communication tools.

Automation in the processing of construction correspondence can prove to be advantageous to its overall financial yield. Automation of correspondence processing can result in lower costs, increased efficiency, timely distribution, and better information handling.

Construction industry involves excessive information being transferred within departments, on a daily basis. Personnel engaged in this field are too much occupied with their duties that if they are simultaneously going through all the documents transferred, and analysing them all, might affect their other critical responsibilities. To help eliminate this serious situation, increasing interaction and networking among diverse industry units can boost productivity and efficiency. Accurate information access is crucial in today’s era for the development of cooperative mechanisms and communication tools.

Construction documents content analysis and processing automation can result in easier data management, rapid responses, more efficient content management. In addition to these benefits, it can reduce uncertainty, reduce information loss, and

reduce work. Majority of the information in fields are present in textual form which could be subjected to content analysis and extraction. As large quantity of data is transferred, it becomes an obstacle in data analysis.

Chapter 02

LITERATURE REVIEW

2.1 GENERAL

During the course of their work, employees often acquire vast amounts of written material. Work papers, agency documents, meeting transcripts, past evaluations, and other sources of information all include relevant data that is difficult to aggregate and evaluate due to its diversity and unstructured nature. A collection of processes for gathering and arranging this data is known as content analysis.

Listing the important topics contained in written information is one technique to begin arranging it so that it may be studied.

Construction is a highly information-intensive sector worldwide, with the success of any project relying heavily on fair access to, effective administration of, and complete analysis of communication or correspondence data. A construction project spans over a long period of time. It may involve on-site production, a big number of personnel, and most crucially, a highly variable firm staff. Furthermore, a tremendous volume of documents containing essential data are produced and exchanged during the project lifecycle, which makes this sector unique.

Textual papers e.g., Letters are the most common method of information transmission in building projects. As a result, effective information gathering of the enclosed content becomes a difficulty depending on the structure of these letters.

2.2 CONTENT ANALYSIS AND INFORMATION EXTRACTION

Content analysis is process carried out to convert immense amount of data from documents shared in fields to smaller portion of required information. The documents shared could be letters, emails, contracts files or any information shared within company.

Information Extraction (IE), or discovering and extracting a sub-sequence from a given series of instances that contains the information we're looking for, is a critical problem with several practical applications. The sub-sequence that we are interested in is discovered and extracted from given data using learnt model(s) using various ways during extraction. On the basis of predetermined metadata, the extracted data is then annotated as prescribed information. (Tang et al 2008).

The web produces an immense amount of useful information that is typically laid out for its users, making it complicated to pull required data from a diverse of sources. To resolve this concern, having authentic, adaptable Information Extraction (IE) systems that shift Web pages into program adjustable structures like a relational database will become critical. Although there are various strategies for extracting data from Web pages, which have been developed, yet there has been almost no effort to compare them. The drawback, here is that the extraction tasks addressed by different tools are not similar, the results provided by multiple tools can only be directly compared in a few conditions.

In these years, text data being shared in communications are immense in quantity, and the methods used to deal with them are outdated, neglected. No worker or construction employ is competent of reading, analyzing and integrate on a regular

basis. Information that has been bypassed as well as lost opportunities—has encouraged people to study into techniques for establishing textual order wilderness. Information retrieval (IR) and information selection are the most frequent processes. Information extraction is a relatively new process introduced into fields.

This method can be operated to convert enormous number of raw materials with an immense volume of potentially useful material into smaller content. With application of this knowledge, an IE system can refine and reduce the given context to a more refined form. IE initiates with a gathering of such messages and then converts them into data that can be analyzed and studied more easily It retrieves relevant information from relevant text segments and isolates relevant text fragments.

The purpose of IE research is to develop systems that discover and link useful data while discarding unnecessary or extraneous data. (Cowie et al 1996)

2.3 CONVENTIONAL METHOD OR MANUAL CONTENT ANALYSIS (MCA) AND INFORMATION EXTRACTION (IE)

The industry's traditional approach of content analysis and information extraction is totally dependent on manual input and construction correspondence management. MCA is defined as a methodical, recurring method for condensing enormous amounts of text into fewer content categories using manual labour and specific coding criteria.

Manual content analysis (MCA) can be defined as the careful scanning and sorting out of relevant information from letters by individuals. Manual content analysis can be defined as the careful scanning and sorting out of relevant data from letters by individuals.

In the construction business, technical data is shared at an astonishing rate, and it is easily recognized as the dominant medium of information transmission. The proper management of information becomes a key challenge based on the structure of document. Due to the vast amount of data, a manual method to building linkages between the information gathered from them and their analysis is unfeasible. The MCA procedure is expensive and results in additional costs. The act of CA is heavily influenced by a person's ability level because knowledge extraction and recovery of valuable information from given structured or unstructured data is a difficult process that demands expertise. Furthermore, the information retrieved is not always constant and varies.

Manual IE also presents a significant challenge when it comes to the analysis of large amounts of data because productivity level varies with time and person, resulting in error-prone extracted output. Furthermore, MCA causes data loss and updates to be slower. As a result, the MCA and IE are processes plagued by a slew of issues. Below are some of the problems associated with the processes of MCA and IE.

The table on the next page, Table-1, enlists all the inefficiencies or in other words disadvantages of manual content analysis of otherwise known and referred to as MCA with references from the previous studies.

Table 1 - Inefficiencies in Manual Content Analysis and Information Extraction

Inefficiencies	**References**
Time-consuming in processing (time inefficiency)	(Kondracki *et al.*, 2002, De Graaf & Van Der Vossen, 2013, Jayaram & Sangeeta, 2017, Karanikas *et al.*, 2000, Stemler, 2001)
Costly	(Kondracki *et al.*, 2002)
The complexity of text retrieval	(M. Lewis & Steedman, 2013, Aslam *et al.*, 2003, S. C. Lewis *et al.*, 2013)
Reduced work efficiency	(Truman, D. B. *et al.*, 1952, Berelson B. *et al.*, 1952)
Subjective approach	(Nunez-Mir *et al.*, 2016)
Lack of consistency	(Allen, 2017, De Graaf & Van Der Vossen, 2013)
Storage problems	(Martínez-Rojas *et al.*, 2015)
Disregards the context	(Nunez-Mir *et al.*, 2016)
Relevancy	(Kondracki *et al.*, 2002)
Data loss	(Martínez-Rojas *et al.*, 2015)
Inadequate data maintenance	(Allen, 2017)

2.4 AUTOMATED CONTENT ANALYSIS (ACA) METHOD INFORMATION EXTRACTION (IE)

The use of various approaches by computers to extract significant patterns and associations from huge textual sources is known as automated content analysis. Text mining is a branch of information technology that deals with automated textual analysis. Text mining can automate the process of information extraction by using specialized procedures.

2.4.1 TEXT MINING

Text mining, also referred to as text data mining, is the procedure of extracting high-quality information from text. It can also be equivalent to text analytics. It requires "the automatic extraction of information from several written resources by a computer to discover new, previously undiscovered information." Websites, books, emails, reviews, and articles are examples of written resources. Statistical pattern learning is commonly used to provide high-quality information by generating patterns and trends. We can differentiate three main and distinct parts of text mining as: information extraction, data mining, and a KDD (Knowledge Discovery in Databases) approach.

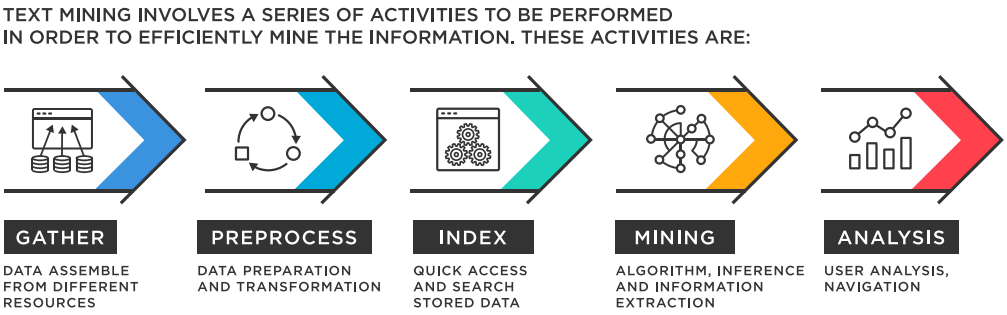


Figure 1 - Process explaining Text Mining

The content induced in the mining process is melded to adjust with the system and analysing, examine and comprehend the processed data, defines text mining in a professional way. In text mining model mining expertise are utilized on keywords that can be any subject or critical entity and these are pulled out from provided context. To execute the process in correct manner, the procedure needs to be well constructed, and subsequently we take use of information extraction mechanism. The overall purpose is to convert raw data into concentrated relevant data for study with the application of natural language processing (NLP), various approaches, and systematic approaches. The interpretation of the acquired data is a crucial part of this procedure.

To educate the system a bunch of natural language samples and either modelling the given samples sets for adjusting with the type of data that would be provided. When it comes to text mining, the document is the most important component. We describe a content as a written data that can be found in a diverse collection.

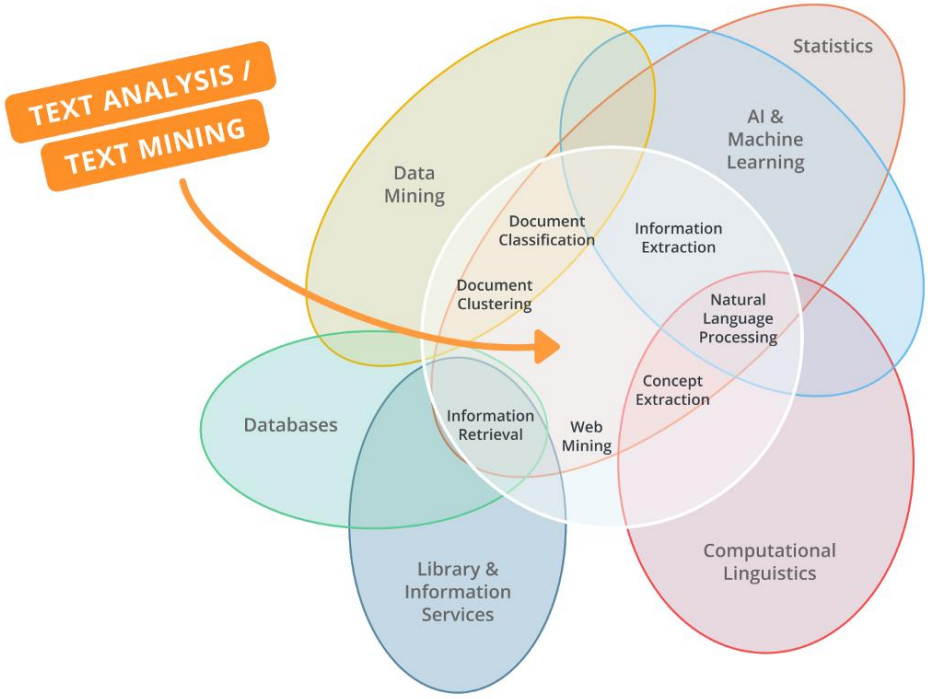


Figure 2 - Venn Diagram showcasing Text Analysis / Text Mining

In current years, carrying out the company forward with less physical manpower is a more intelligent decision. To accomplish so, you'll need some advanced software to help you diminish the difficulties of manual work. In a company, the majority of physical labor is focused on written context. With document capture software, manual data entry, which is prone to errors, may now be automated. In any paper-based office, this is quite advantages.

Manual content analysis methods are not built to carry out the extraction of large quantity of media texts. Instead, a different method for example a automated solution that would pull out the relevant data from the document and compile it in such a way that would allow us to do a content analysis in terms of these challenges and inefficacies. For history it is clear that computer programs are advanced at

recognizing and separating information based on formational characteristics and organized patterns, and can follow rigid guidelines with perfect accuracy, an automated approach was well adjusted and appropriate.

If an individual starts to analyze all data and extract it, it would have taken far too long and resulted in errors. To assist us in completing this task, we wrote a Python code. Python was selected for process as it is a basic yet efficient advance normal use programming language that is well-adjusted to writing scripts for parsing vast amounts of consistent data.

The restriction and possibilities of large information for excessive, quantitative content analysis in communication study have been described in this article. The capacity of traditional content analysis has been strained as the movement from large communication to connection communication has resulted in massive bodies of open communication online, pushing online information analysts to search for modern ways to the persistent complications of choosing and coding data. Automation methods provide up new methods for resolving such problems, but they have limitations in terms of what they can gain on their own. As a result, we believe that in many circumstances, a collaborated methodology that combines computational and manual methods throughout the content analysis process is optimal for researchers.

For traditional content analysis of massive data being shared, there are trade-offs when applying manual content analysis to Big Data—either a drop in sample size or an increase in human coders. If we were presented with a lot of data and needed to reduce it to a size it would be a hassle. If compared with orthodox methodologies,

we can use computational tools to magnify the work of a small group of coders, allowing them to figure out the data more rapidly with minute details. (Lewis et al 2013)

ACA, unlike manual methods of literary synthesis, can generate large volumes of material in significantly shorter time frame while removing human unfair judgment. Because of its general productivity and long-term advantages, this method can be used for a diverse range of already published content including both exploratory and systematic reviews aimed at answering more specific research questions. ACA has the potential to fill a significant methodological gap and hence assist to the advancement of ecological and evolutionary research by allowing for more extensive and complete assessments of large amounts of material.

According to the trend in which new information is entering market and all these content needs to be dealt, for that the capabilities of researchers have to be more advanced to properly handle all these communications and information, and otherwise the issues related to them would grow rapidly.

Automated information removal is a set of procedures that use statistical model, such as theme models to uncover a body of literature's latent subject composition. The purpose of these algorithms is to recognize certain themes and categorize literature based on their presence.

Automated content analysis has two essential qualities that make it useful for reviewing and synthesizing large amounts of data. For starters, ACA can handle enormous amounts of literature far faster than manual methods. To show practical example of the tool's processing capabilities, the python-based ACAtool, Gensim,

was used to process all articles on It was discovered that the tool could process 16 000 documents per minute. Other than the amount of literature available within the scope of the inquiry, there is no theoretical limit to the amount of text that can be analyzed by ACA techniques. The ability to analyze vast amounts of text quickly enables for the study, re-analysis, and synthesis of far bigger samples of literature. Unintentional human bias is the second attribute of ACA that contributes to its value. Human classification is impacted by a variety of factors (such as weariness, personal biasness, and state of mind), many of which factors are ignorant of and hence unable to disclose. ACA can help to deal these factors, which could help to limit subjective human bias. “ACA builds concept categories from text data using methodologies based on ‘grounded theory' – the reciprocal informing and shaping of data collection and data analysis through an emergent iterative process.” (Nunez-Mir et al 2016)

2.4.2 NATURAL LANGUAGES AND COMPUTING

There has always been requirement of interacting with computers, and to interact their must exist a communication channel. In case of computers, it has been programming languages. But these programming languages are not for everyone to understand. Human languages and computing languages are so different that not only on syntax or structure but most importantly the choice of words. There has been a constant development of new, easy to use languages for humans to code in. In the past, the assembly language was used which didn’t make sense to normal human, as the time progressed and development continued more user-friendly programming languages were introduced which included human understandable

words, but then the syntax was still an issue. In these contemporary times, we have some very easy to learn programming languages which have simple syntax and no more complex naming like Java.

Aside from the programming languages getting easy to code in, there has been significant development of Artificial Intelligence (AI), this technology is still somewhat limited in use but it is constantly being improved and nowadays AI can understand several basic commands in natural or human languages. It will take quite more time till it can understand more complex human language. But when it will be complete, we would be able to interact with the computer in the natural or human language, just like we interact with each other.

The specific area of AI which focuses on natural or human languages is called NLP (Natural Language Processing).

2.4.2.1 Natural Language Processing (NLP)

With the advent of computer aroused the need of interacting with it and one way of interacting with it, which is most easy one is natural languages. But this is a Work In Progress (WIP). NLP is the process for making computers understand what we want in human languages not in the programming languages.

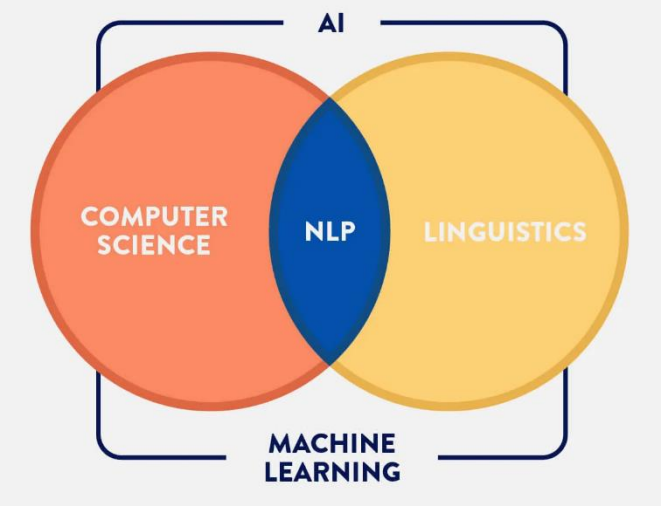


Figure 3 - Venn Diagram of NLP

2.4.2.1.1 Natural Language Understanding (NLU)

It is the use of data, human language texts, and making computer to understand how the natural languages work. For that purpose, the sentences in human languages are split in parts and how they relate to each other. For NLP to work NLU is the integral part of making the AI to understand the human languages.

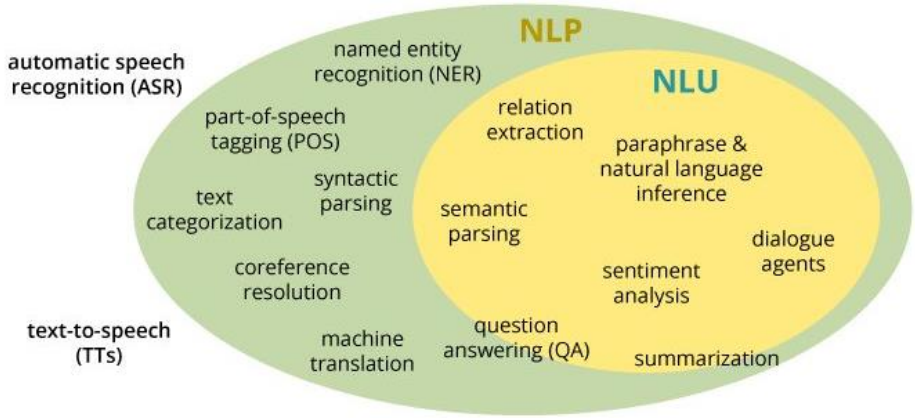


Figure 4 - Venn Diagram for NLP and NLU

To make AI understand the human languages as we speak and write, NLU framework was developed, as most of the time humans don't usually follow all the ideal rules of grammar and sentence structure and sometimes there are similes and metaphors to make the language more complex for computer to understand. Therefore, the need to develop the framework which can train and assist the AI in understanding of human languages was developed, namely NLU.

Natural Language Understanding (NLU) is a mix of several components each serving its own purpose. These seven parts are namely:

2.4.2.1.1.1 Lemmatization

It is the process of reducing the inflected forms of a word into a single form for AI to perform analysis easily.

2.4.2.1.1.2 Stemming

It is to cut the inflected words to their root form.

2.4.2.1.1.3 Morphological Segmentation

In this the words are divided into morphemes, which can be described as the smallest and meaningful unit or lexical item.

2.4.2.1.1.4 Word Segmentation

Continuous texts are divided into distinct units in this process.

2.4.2.1.1.5 Parsing

To perform grammatical analysis on sentences and understanding its sentence structure.

2.4.2.1.1.6 Part-of-speech (POS) Tagging

As the name implies, it identifies the Parts of Speech (POS) for each and every word in the sentence.

2.4.2.1.1.7 Sentence Breaking

This step breaks the continuous texts in sentences by placing sentence boundaries. This makes the Continuous text easy for the AI to process on.

2.4.2.1.2 Natural Language Generation (NLG)

It is the opposite of NLU; NLU is the making of AI to understand the human languages, while NLG is the process of making the AI generate a human understandable text. It consists of following tasks:

2.4.2.1.2.1 Text Planning

It is done to fetch the content which is relevant from the existing knowledge base.

2.4.2.1.2.2 Sentence Planning

It involves the process of selecting the appropriate and proper words, generation of sentence that is meaningful, and to set the tone of the sentence accordingly.

2.4.2.1.2.3 Text Realization

This is the final step to form from the sentence structure from the sentence plan generated as a result of the above processes.

Chapter 03

RESEARCH METHODOLOGY

This section defines the procedures adopted to achieve the desired output, of automated content analysis, including but not limited to the pre-liminary research, identification of gap and utilizing the new technology in making the process automated and a lot smoother.

3.1 METHODOLOGY

3.1.1 RESEARCH DESIGN

The preliminary study which included finding of gap in the previous studies, which assisted in formulating the problem statement and selection of research objectives. After that the inefficiencies in the MCA were dealt with the help of program, which is a rule-based NLP framework. Below is the detailed Program working.

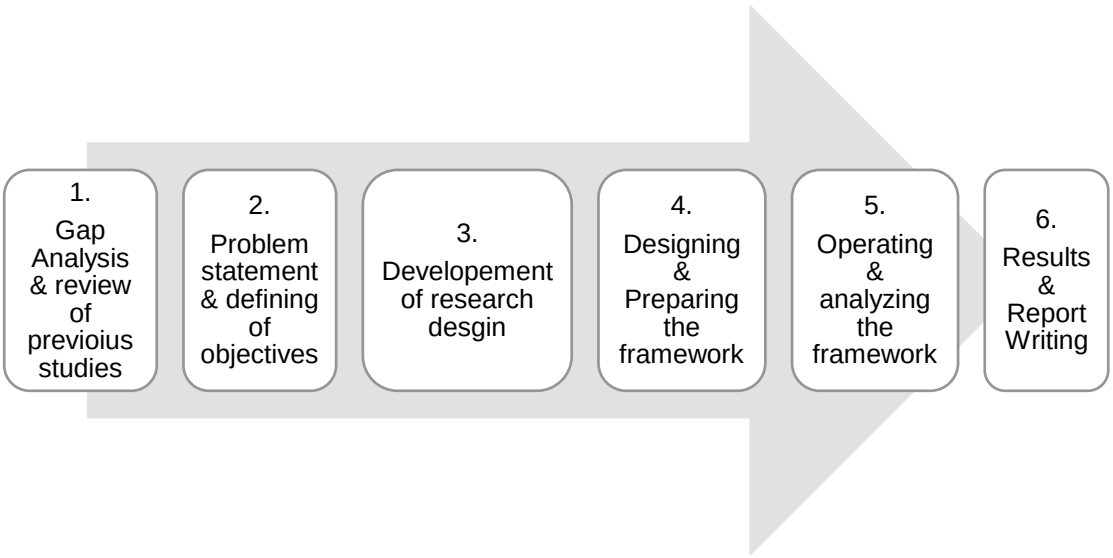


Figure 5 - Research Design Flow Diagram

3.1.2 PROGRAM INITIALIZATION

When the program is started, first the things you need a Compiler/Interpreter. The core components include libraries, variables, and required files or data. These components are listed below:

3.1.2.1 LIBRARIES

In programming terminology, a library is a permanent resource utilized by computer programs, often for software initialization, development, and execution. These may contain pre-written/pre-compiled code and functions, classes, values, or type specifications. A library is a previously written set of commands, that can assist us in calling certain functions eliminating the need to code again and again for simple purposes. It is once done that for the easy use and integration in future. Different libraries will have different functionalities and have certain restrictions on usage, but libraries are nothing on their own. When the program is compiled the thing that is integrated are libraries including but not limited to:

3.1.2.1.1 doc2txt

This python library extracts the contents like text and other smart arts from a Microsoft Word document file (extension: <filename>.docx).

3.1.2.1.2 re

This module provides regular expression matching operations for the 8-byte based strings and Unicode strings. As both patterns have different structure they cannot be compared normally.

3.1.2.1.3 os

This library enables the program to interact with the host operating system, especially in case or reading and writing a file. This modulea also enables to manipulate paths of files. It is also helpful in creating temporary files and directories.

3.1.2.1.4 glob

The python-based module finds all the path-names matching a specified pattern in accordance to the rules used by the shell.

3.1.2.1.5 datetime

Date and Time are passed to programs through this module. It also provides Date and Time in several different styles and formats.

3.1.2.1.6 dateutil.parser (parse)

This module is used to parse a date and/or time passed to it. This module also tries to make sense out of unlikely or non-conventional input formats.

3.1.2.1.7 numpy

NumPy is the fundamental package for any array computation done in the Python. It provides: N-dimensional array, complicated mathematical functions, tools for integrating code from C/C++ and Fortran, useful linear algebra, Fourier transform, random number generation capabilities, and a lot more. Besides it being a scientific

tool, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary datatypes can be defined. NumPy can also integrate with a wide variety of databases easily, effortlessly and quickly.

3.1.2.1.8 csv

The so-called CSV (Comma Separated Values) format is the most common import and export format for spreadsheets and databases. CSV format was used for many years prior to attempts to describe the format in a standardized way in RFC 4180. The lack of a well-defined standard means that subtle differences often exist in the data produced and consumed by different applications. The csv module implements classes to read and write tabular data in CSV format and solve this lack of standardized structure.

3.1.2.1.9 argparse

This module makes it easy to build user friendly Command Line Interface (CLI). The program identifies what arguments it requires, and it will decide how to parse those. This libarary also automatically produces assistance and usage prompts and issue errors when invalid arguments are passed to it.

3.1.2.1.10 os.path & os (path)

This library implements some valuable functions on path-names. To read & write files, or to access the file-system. The path as parameters can be passed to it as strings, or bytes, or any object applying the os.PathLike protocol.

3.1.2.1.11 openpyxl (load_workbook)

It is a Python library to read, write, format, and perform some other operations on Microsoft Excel xlsx/xlsm/xltx/xltm files.

3.1.2.1.12xlsxwrite

This Python module is for populating files in the Microsoft Excel .xlsx file format. The data that can be populated include text, numbers, formulas and hyperlinks. It can also be used to multiple worksheets, and it supports features such as formatting and many more.

3.1.2.1.13pandas

It is a Python package that provides quick, supple, and communicative data structures with the goal to make working with "interactive" or "characterized" data both effortless and spontaneous.

3.1.2.1.14sys

This module provides access to some variables used or maintained by the interpreter and to functions that interact strongly with the interpreter.

3.1.2.1.15subprocess (check_output)

This python module gives the ability to issue new processes, link to their input/output/error stream, and acquire their return outputs. It also intends to substitute numerous older modules and functions: os.system, os.spawn*.

3.1.2.1.16shutil

This library offers numerous high-level processes on files and groups of files. Functions are provided which support file replication and deletion.

3.1.2.2 VARIABLES

The data in programs are stored in two ways – it can either be in form of user input or locally stored within programs. The data that is stored inside the program is called

Variables. Each variable is unique and have a unique identifier, variables are also defined on the base of their data-types, and a value corresponding to its data-type assigned to it. The value assigned that variable can change as the program executes, therefore, the name Variable.

This data can be unknown or known based on the assignment of value to the variables. Variables can be thought of as ‘containers’ which can hold more than one value. Their only purpose is to store and label data in the memory. After that this variable can be recalled anytime in the program using its unique identifiers.

There are two types of variables in a program, local and global variables. A global variable has a global scope, meaning that it is accessible and usable all over the program. Global variables are mostly non-value changing variables, whose extent is the whole runtime of the program. The Global variables are introduced and assigned values respectively in the program, to assist in the program functionality.

3.1.2.3 DATA

Data can be defined as a group/cluster of discrete facts, and statistics; often numeric. The data can also be said to be a set of values of quantitative or qualitative variables about several persons or objects, while a datum is a single variable bearing a single value.

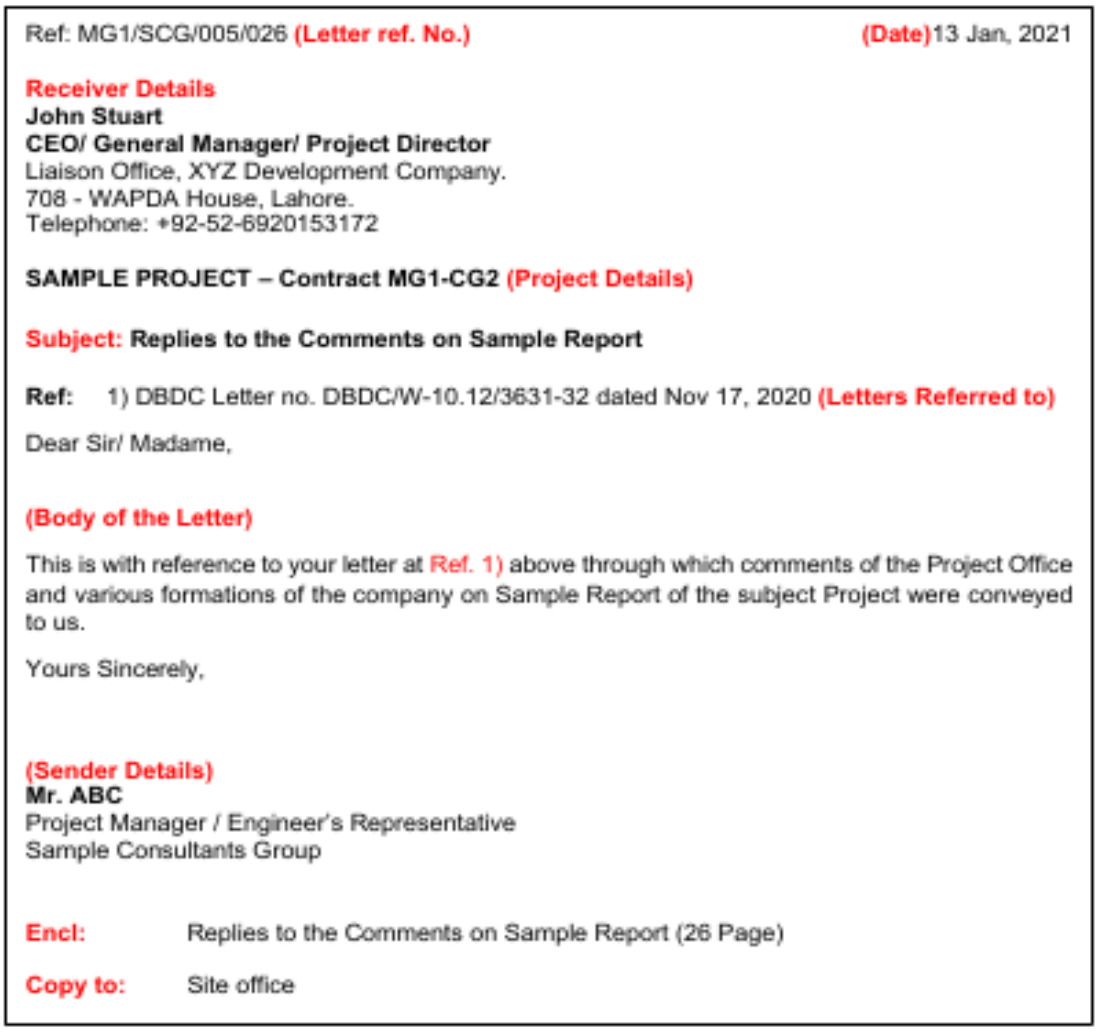
Till now the preparations to run main program were being integrated and implemented. Now the data will be integrated in the process. As they are done the main program is initialized.

3.1.2.3.1 DATA PREPARATION

There are several checks even before the main program even starts the desired process on files, these checks include the input directory of files and the other files necessary for program to run, in our case:

3.1.2.3.1.1 Letters

These are the input data set for our program to work on and generate a summary of them. These are stored in a directory which will be passed to the program in the full path form as an argument. These will be in a directory named “data_to_process”, after processing these will then automatically be moved to the directory named “processed”, to signify that these letters/files have been processed and their information which could be extracted is filled in the summary sheet/file.



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 6 - Sample Format of Letter

3.1.2.3.1.2 Summary Sheet

This file contains the summary of the letters, its columns in which the data will be populated are given in the order and the file template is ready for program to fill in with the data that it will extract from the source data, in our case, letters.

3.1.2.3.1.3 Purpose Database File

This file is a database file in the comma separated values format, or commonly referred to as a csv file. This file contains all the keywords to be used in the program, and as the name suggests, the keyword regarding the purpose. In our case, it includes the certain significant keywords regarding the construction industry including but

not limited to construction works, related to schedule and costing. This file can be updated to include more keywords, without any requirement to update the code.

3.1.2.3.1.4 Designations Database File

This file in the simple text format, which is a universal text format, bearing an extension of .txt. This inter-operable file contains the designation or simple the title of post, which can help in the extraction of the designation of the sender and receiver. This file is also easy to update, just add new titles in the new lines of the text file, without any requirement to update the code.

3.1.2.3.2 DATA AVAILABILITY

If all are present the program will continue otherwise terminate with error even if one of the above-mentioned things is not present with an error showing an error message describing what is missing.

A detailed chart showcasing the use of program working to fetch files/data files for processing, and how the program logically decides to proceed further by determining whether all the abovementioned files are present where the program needs them to be. This flow chart id depicted in Figure-7.

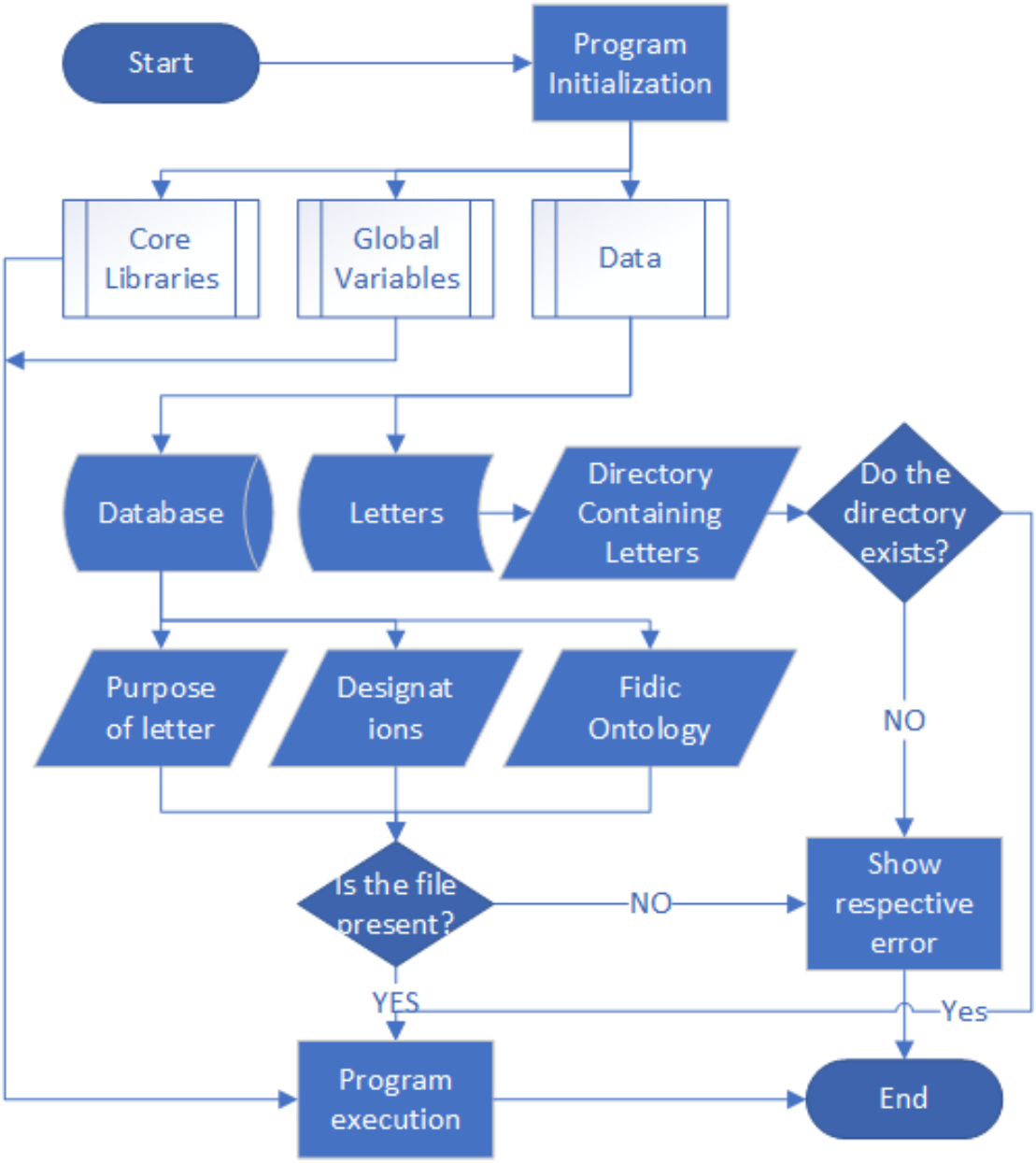


Figure 7 - Flowchart of File Validation Process

3.1.3 PROGRAM EXECUTION

Each program runs in a certain order, the in-order steps of how this program works and how the analysis of letters is done using sequence and output is generated.

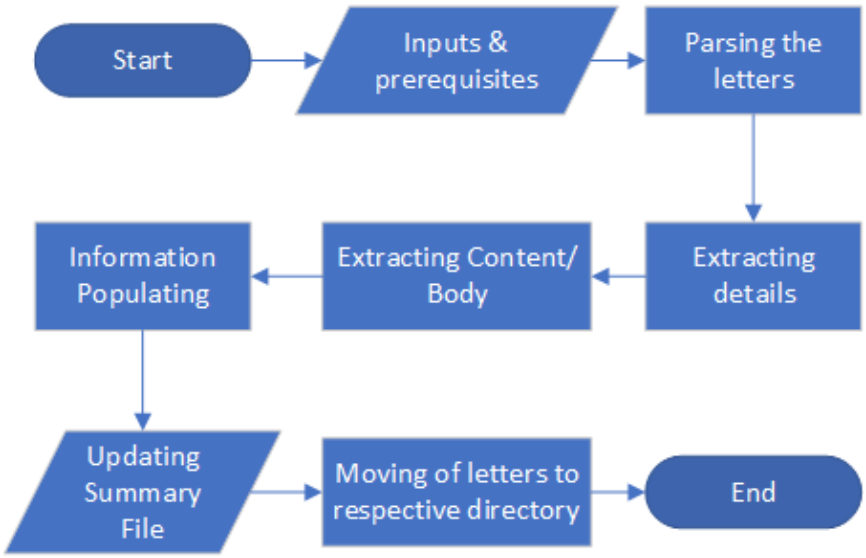


Figure 8 - Brief Flowchart of Program Working

3.1.3.1 DATA ANALYSIS

Now before the analysis is done, the program needs to extract data from database files and save them in respective array. The purpose phrase array is cleared to remove redundant or garbage entries then the contents are changed to floats, then the program form a list using those processed floats and saved in purpose phrase array. The designations array is cleared to remove redundant or garbage entries then the contents are changed to floats, then the function form a list using those processed floats and saved in designations array.

When these global arrays are updated and the program reaches its first stage of the analysis, which is the forwarding of the files (directory of letters and summary file) as arguments to the processing functions.

The first step is to process the summary file to extract the features which is done by parse extract. To the parse extract function the program forward the data path, path

to letters, and summary file. Which then appends all the documents files or letters together, then the counting for the number of letters is done by using counting variables including but not limited to:

3.1.3.1.1 Local Variables

3.1.3.1.1.1 cnt_deadline

This variable will contain the deadline related to the specific letter, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.2 cnt_exceeded

This variable will contain the value of any mention of date exceeding related to the specific letter, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.3 cnt_letters

This variable will contain the reference number related to the specific letter, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.4 cnt_addressed

This variable will contain the status of reply related to the letter, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.5 cnt_cost

This variable will contain the cost or monetary value mentioned in the specific letter, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.6 cnt_timeframe

This variable will contain the detail of the time frame related to the specific letter, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.7 cnt_ref

This variable will contain the references of past/previous letters provided in the specific letter, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.8 cnt_encl

This variable will contain the enclosure details of the letters, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.9 cnt_copyto

This variable will contain the information regarding to who else is this copy of letter sent to, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.10 cnt_clauses

This variable will contain the any mention of clauses, sub-clauses, articles or sub-articles, if any, present in the letter, when initialized this variable will be assigned

the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.11 err_files

This variable will contain the information related to the specific letter(s) in this any error made the program to discard the letter, when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.1.12 cur_rec

This variable will assist the program in populating the fields of table of the table in summary file by keeping the record of which file number data is being filled in. when initialized this variable will be assigned the value of 0 or “zero” or empty to avoid and garbage value from being populated in the variable.

3.1.3.1.2 FUNCTIONS

3.1.3.1.2.1 Parsing

After initializing libraries and variables, the program then proceeds to process file content and save it then separating these contents with a new line after splitting lines, it converts all lines into list. Then it initializes the column variables which will help us in storing the data from array into the column of summary file it matches using the while loop. This loop works in an iterative manner for each line, starting from the first one to the last one.

3.1.3.1.2.2 Extraction of Details

The first thing to be extracted from the array is designation, as they usually come in top portion, it is assumed that it will be stated in the first 20 lines, this many lines so that the program doesn't miss it. It retrieves Receiver, Receiver's Designation, Receiver's Organization. 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.

Copy to, 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 sender details would to be extracted from the data/letter, including sender, sender's designation, and organization of sender; and trying to locate it before the letter ending, which usually starts with the key word "yours sincerely".

3.1.3.1.2.3 Extraction of the data from body of letter

Clause, Sub-Clause, Article if there are any mentioned, hold important place and therefore they are extracted before anything else. As the presence of any such thing may mean to refer to clause for something important or change in the terms.

Then it proceeds to find any deadline in the body, and if so then it tries to remove redundant words like by, till, before. The remove the number's suffixes like "st" in First (1st), "nd" in Second (2nd), "rd" in Third (3rd), and "th" in number-th (nth). And the additional punctuation is removed from the date i.e. "," and " ". After that the date is parsed in the format for program to understand and input into the excel file. It also helps in determining whether the deadline is extended or not, by using the parsed date and adding the delay resulting in the new deadline. Extension of time is also processed by this program using keyword tokenization of days etc. The program applies the check to look for the keyword "delay", if present what it is followed by for example "by", "till" et cetera.

The program also then extracts the numerical value of currency/cost/amount. Using the keyword extraction rules and currency symbols. The classification on whether the reply to status to the letter is pending or not is also done and updated.

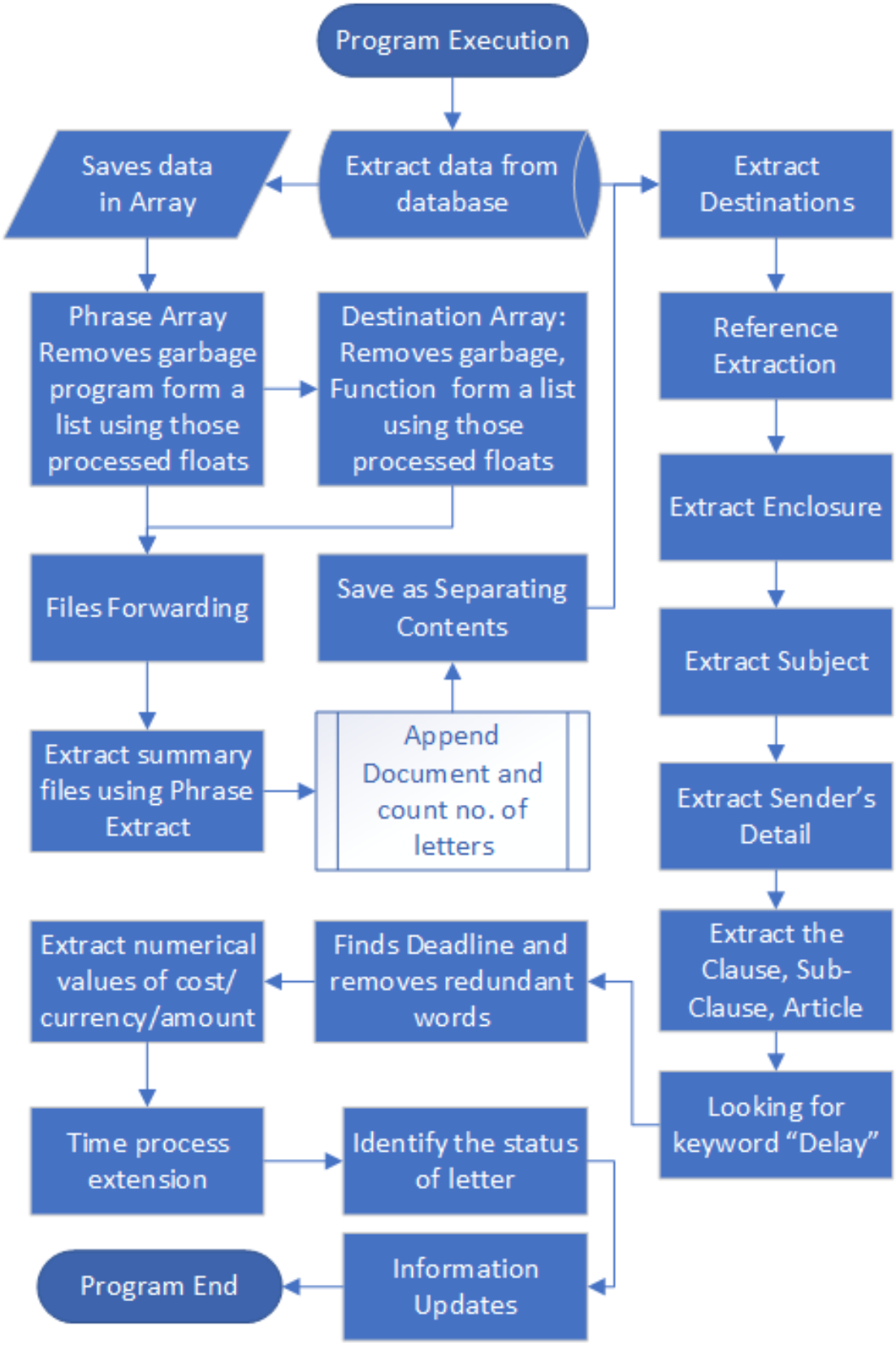


Figure 9 - Information Extraction workflow diagram

3.1.3.2 INFORMATION UPDATING

After that all the information extraction is done, now it's time for data entry for which line is made with all columns to be written into Results csv file. Which includes:

3.1.3.2.1 Aggregated Values and Information

Reference No. of letter	Date of letter
Sender	Designation of Sender
Organization of Sender	Receiver
Designation of Receiver	Organization of Receiver
Project name	Subject of letter
Purpose of letter	Clause and Sub-Clauses
Extension of time	Deadline
Day or mention of time	Cost
Encl	Copy To
Status of reply	Total words

When all the data in the source files are compiled in comma separate values format the process of updating the summary file begins and data in entered the respective columns. If any problem occurs, then the file is discarded with a message of error in file and is moved to the error files directory. Also, to update the status of previous letters which are present in current reference is done by this program and it updated the value of status from pending to addressed.

3.1.4 PROGRAM FINISHING

When all the processing is done, the files are moved to processed directory. And a message is displayed which shows how many files are successfully processed and how many couldn't be processed due to errors, if any. The program is closed afterwards, summary file is updated and now it can be viewed. This excel sheet is our summary sheet, which can be used to locate the exact information we would be needing instead of going for all the letters.

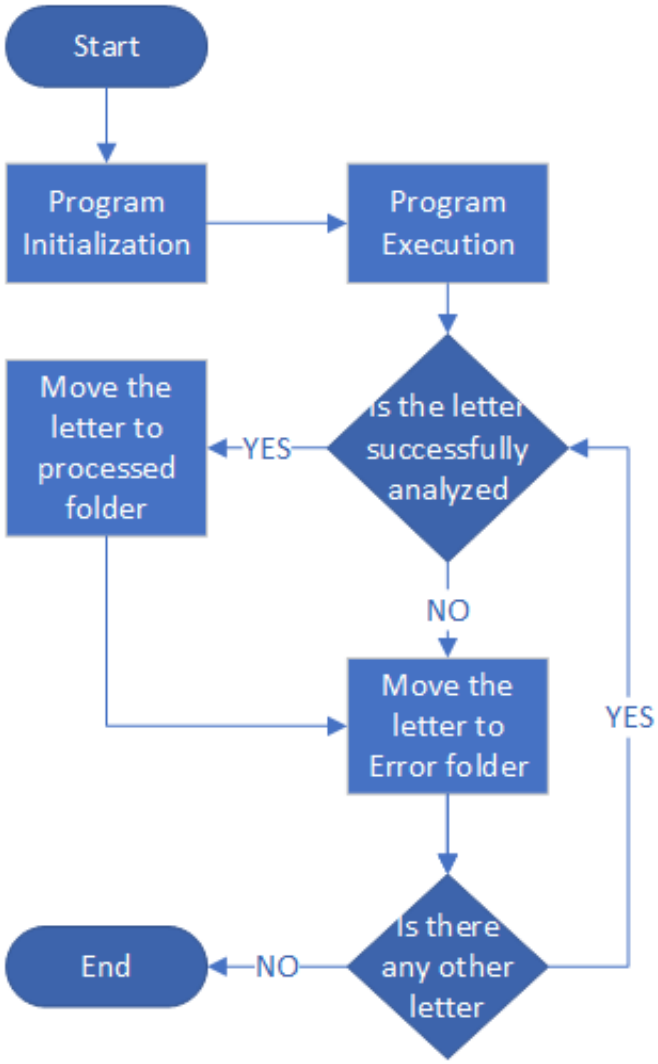


Figure 10 - Flowchart Explaining how letters are moved after processing

Chapter 04

CONCLUSION

Professionals involved in construction industry especially, get influx of letters all the time so a platform was much needed to smoothen up the processing time and effort. Alongside this there were multiple issues attached with manual information extraction which were key factors in delaying the communication, causing errors, loss of information and increased costs. This program automatically analyses the letters and generate summary which it develops/generates. Moreover, it was manually checked for the effectiveness of program by comparing the output summary sheet and the letters. Now that this research has been incorporated in current construction industry, information transfer would become trouble free. Basically, it bridges the gap between an unpractical approached methods with a more profound method.

FUTURE RESEARCH

Today's era is more in practice with online communication and most of the correspondence that happens is preferably done through online means, so applying this information extraction process on online means is recommended, and as the documents are nowadays being shared in soft copy form. Below are some of the recommendations for future research:

- As the online communications takes over, people use mixed languages, especially in our case local languages, so a framework to analyse those local languages to extract information would be a good addition.
- Image based text process can also be focused on extract textual data from scanned pictures of official documents, screenshot of web-sites, or any image with a few characters. And then analysing those letters using this program.
- In the field of AI, Optical Character Recognition (OCR) technology can be used to extract information from hard copies and then converted to scanned image format which can then finally fed into the system.
- Emails are nowadays treated as official means, so a program to analyse information in emails would be great.

Chapter 05

REFERENCES

Allen, M. (2017). The SAGE Encyclopedia of Communication Research Methods. In *The SAGE Encyclopedia of Communication Research Methods*. https://doi.org/10.4135/9781483381411

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., ... & Zhai, C. (2003, April). Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst, september 2002. In *ACM SIGIR Forum* (Vol. 37, No. 1, pp. 31-47). New York, NY, USA: ACM.

De Graaf, R., & 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). https://doi.org/10.1515/commun-2013-0025

Jayaram, K., & Sangeeta, K. (2017). A review: Information extraction techniques from research papers. *IEEE International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2017 - Proceedings*. https://doi.org/10.1109/ICIMIA.2017.7975532

Karanikas, H., Tjortjis, C., & Theodoulidis, B. (2000). An Approach to Text Mining using Information Extraction. *Proc. PKDD 2000 Workshop on Knowledge Management Theory & Applications*, *Dm*.

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

Lewis, M., & Steedman, M. (2013). Unsupervised induction of cross-lingual semantic relations. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting and Electronic Media*, *57*(1). https://doi.org/10.1080/08838151.2012.761702

Martínez-Rojas, M., Marín, N., & Vila, M. A. (2015). An Approach for the Automatic Classification of Work Descriptions in Construction Projects. *Computer-Aided Civil and Infrastructure Engineering*, *30*(12). https://doi.org/10.1111/mice.12179

Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N., & Fei, S. (2016). Automated content analysis: addressing the big literature challenge in ecology and evolution. In *Methods in Ecology and Evolution* (Vol. 7, Issue 11). https://doi.org/10.1111/2041-210X.12602

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research and Evaluation*, *7*(17). https://doi.org/10.1362/146934703771910080

Truman, D. B. (1952). Content Analysis in Communications Research. By Bernard Berelson.(Glencoe, Ill.: Free Press. 1952. Pp. 220. $1.25.). *American Political Science Review*, *46*(3), 869-873.