國立臺灣大學工學院土木工程學系
碩士論文
Department of Civil Engineering

College of Engineering

National Taiwan University

Master’s Thesis

以自然語言處理方法自動識別施工排程中的危害

Automated Hazard Identification in Construction Scheduling

謝文龍

Edwin Shiady

指導教授：謝尚賢博士

Advisor: Shang-Hsien Hsieh Ph.D.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202401595

國立臺灣大學碩士學位論文
口試委員會審定書

NATIONAL TAIWAN UNIVERSITY
MASTER'S THESIS ACCEPTANCE CERTIFICATE

建築施工進度中的自動化危害識別：自然語言處理方法

Automated Hazard Identification in Construction Scheduling: A Natural Language Processing Approach

本論文係　謝文龍 (R11521618)　在國立臺灣大學土木工程學系電腦輔助工程組完成之碩士學位論文，於民國113年07月03日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Civil Engineering, Computer-Aided Engineering Division on July 3, 2024 have examined a Master's Thesis entitled above presented by Edwin Shiady 謝文龍 (R11521618) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

謝尚賢
（指導教授 Advisor）

林耕宇

紀乃文

系主管Director： 葛宇甯

doi:10.6342/NTU202401595

摘要

建築業多年來一直是工安死亡事故居高的行業之一。儘管企業和勞動部職業安全衛生署（Occupational Safety and Health Administration, OSHA）努力降低事故發生率，由OSHA記錄的違規案件仍居高不下。良好的安全規劃，尤其是在專案早期階段，是防止未來事故的必要條件。為了實現這一目標，多年來人們進行了大量的研究， 包括電腦視覺、建築資訊建模（ Building Information Modeling, BIM）、規則化編程和自然語言處理（Natural Language Processing, NLP）的應用。這個研究的目的在於為施工排程建立一個危害識別系統，以便在專案早期階段識別危害。本研究方法選擇了詞頻-逆文檔頻率（Term Frequency – Inverse Document Frequency, TF-IDF）方法，並結合關鍵詞的映射，以創建一個能夠識別危害類型、頻率和來源的模型。透過從排程中提取關鍵詞並將其作為搜尋OSHA數據庫的輸入詞，TF-IDF能夠在事故的最終敘述中搜索到相關危害記錄。根據模型在訓練和測試過程所獲得的閾值，最終敘述被篩選出來。總體來說，訓練和測試顯示的正向結果表明TF-IDF能夠在不犧牲精度的前提下展示危害的類型和來源。這項研究將有助於更快速和精確的危害識別，並可作為進一步危害分析的基礎。

關鍵詞：危害識別、危害頻率、危害來源、安全規劃、自然語言處理、詞頻-逆文檔頻率（TF-IDF）

doi:10.6342/NTU202401595

Abstract

The construction industry is one of the industries that has contributed to a high number of work fatalities over the years. There have been numerous attempts to lower the number of accidents either by companies or Occupational Safety and Health Administration (OSHA). However, despite all the efforts to lower the number of casualties, the number of violations cited by OSHA is still high. Good safety planning is necessary, especially in the early stages of the project to prevent future accidents. To achieve this, much research has been done over the years, using technologies that range from computer vision, building information modeling (BIM), rule-based programming, and NLP. This research aims to create a hazard identification system based on a construction schedule so that the hazards can be identified in the early stages of the project by using NLP. The method chosen for this research is TF-IDF combined with mapping of the keywords in order to create a prototype that is able to identify the type of hazards, frequency of hazards, and source of hazards. By extracting the keywords from the schedule and using them as input in the OSHA Database, TF-IDF managed to search through the Final Narrative of accidents to find relevant hazards. The final narratives are then filtered out based on the threshold obtained from the training and testing process. Overall, the training and testing results show positively that TF-IDF is capable of showcasing types and sources of hazards without sacrificing the precision of the results. This research contributes to faster and more precise hazard identification that can later be used as a basis for further hazard analysis.

Keywords: Hazard Identification, Frequency of Hazard, Source of Hazard, Safety Planning, Natural Language Processing (NLP), Term Frequency-Inverse Document Frequency (TF-IDF)

doi:10.6342/NTU202401595

Table of Contents

doi:10.6342/NTU202401595

doi:10.6342/NTU202401595

List of Figures

doi:10.6342/NTU202401595

doi:10.6342/NTU202401595

List of Tables

doi:10.6342/NTU202401595

Chapter 1 Introduction

1.1 Background and Motivation

Every year, the construction industry contributes to the high number of fatalities. Construction safety status from multiple countries can be seen in Table 1.1

Table 1.1: Construction Safety Status in Different Countries

Countries	**Construction Safety Status**
United States	In 2021, nearly one in five workplace deaths occurred in the construction industry. The construction industry accounted for 46.2 (with a total of 1,015 accidents) percent of all fatal falls, slips, and trips in 2021 [1].
China	From 2010 to 2019, there were 6005 fatal accidents causing 7275 in China's construction industry [2].
Taiwan	According to OSHA Taiwan's 2022 Annual Report, construction and construction equipment is responsible for 7.4% of the injuries happened. The number of injuries reported in 2022 was 6,956 accounting illnesses, disability, and death [3].
Japan	In 2021, number of construction fatal accidents recorded by Japan Industrial Safety & Health Association (JISHA) is 288 which contributed to 33% of total working accidents in Japan [4].

doi:10.6342/NTU202401595

As seen in Table 1.1, even though the statistics were taken from four different countries, it can be seen that the construction industry is still one of the most dangerous industries. According to the Occupational Safety and Health Administration (OSHA) from United States Department of Labor, there are four causes for construction fatalities: fall, struck by, caught in-between and electrocutions [5].

Table 1.2 Types of Accidents and Common Causes

Types of Accident	**Common Causes**
Fall	Unprotected sides, wall openings, and floor holes; improper scaffold construction; unguarded protruding steel rebars; misuse of portable ladders
Struck by	Vehicles (mostly trucks or cranes), falling/flying objects, constructing masonry walls
Caught in-between	Bad enter and exit gateway, cave-in, material placement
Electrocutions	Contact with power lines, lack of ground-fault protection, missing or discontinued path to ground, misuse of equipment, improper use of extension and flexible cords

Table 1.2 shows the common causes for the top four types of accidents, in which most of these accidents' causes can be minimized with proper planning and monitoring. However, despite efforts from OSHA to minimize the number of fatalities in construction industries, many violations were found from October 2022 to September 2023. According to OSHA's NAICS Code 236 about construction buildings, the number of violations throughout the inspection period reached 1,930 violations with a total of $8,859,655 penalties [6]. These violations' penalties are not only expensive but can also cause project delays due to project pausing or even complete stoppage by the government. If an accident does happen, the cost of an accident can also be more than investing in safety measures.

doi:10.6342/NTU202401595

1.2 Research Objectives

Safety planning is a crucial part of the project. Without proper safety planning in the project, unwanted circumstances may happen.

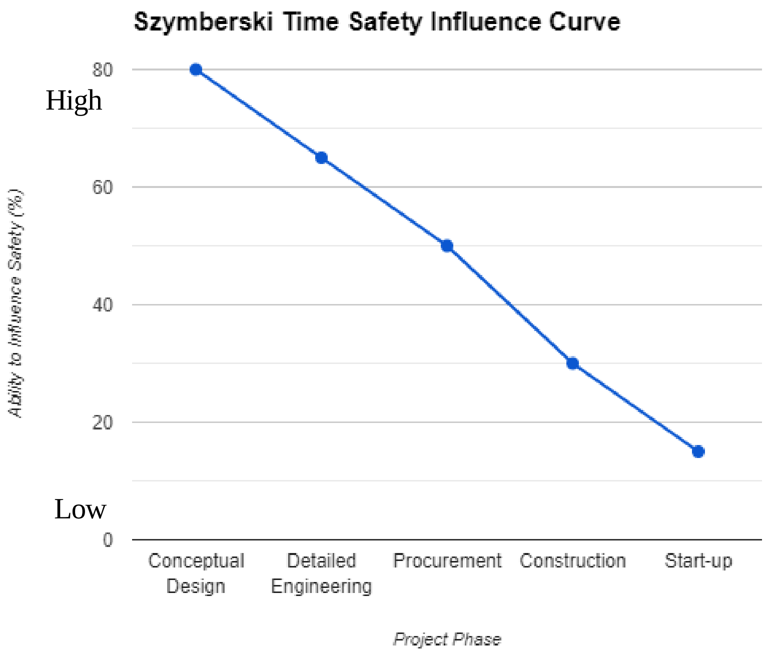

Figure 1.1 Ability to Influence Safety as the Project Begins [7]

Figure 1.1 shows that as the project moves on to the execution phase, the ability to influence the safety condition of a project also decreases. This happens because safety monitoring, while also playing an important role in a project, is unable to negate the safety hazards completely. This happens because sometimes the human responses and reactions are also not quick enough to prevent injuries or even fatalities, which is why safety planning plays an important role to ensure the project is able to proceed smoothly with minimum to no accidents.

Aside from that reason, design-related issues also play a major role in construction accidents. “Thirty-seven (37%) of the 210 workplace deaths definitely or probably had

doi:10.6342/NTU202401595

design-related issues. This happens because of the lack of information about the incident circumstances that ultimately ended up causing difficulties in determining design issues." [8]. Based on that statement, it can be concluded that it is important to have a better understanding of incident circumstances in order to be able to create a better design and safety measurements.

According to the United States Department of Labor, Occupational Safety and Health Association (OSHA), there are several things that need to be noted during a Job Hazard Analysis (JHA) process [9]:

- Involving your employees. This action needs to be done in order to make the employees have a unique understanding of the job.
- Review your accident history. By reviewing history of accidents and occupational illnesses that needed treatment, losses that required repair or replacement, and any near misses, stakeholders can understand events that will be indicators that the existing hazard controls may or may not be adequate to the project's needs.
- Conduct a preliminary job review. Discussing with employees about hazards that they know exist in their current work and surroundings.
- List, rank, and set priorities for hazardous jobs. Listing out jobs with hazards that may present unacceptable risks based on those most likely to occur and with the most severe consequences.
- Outline steps or tasks. By analyzing the steps of a job, information can be broken down to a detailed level that makes hazard analysis better.

Other than those criteria, the OSHA guidelines also mention 5 important questions to ask when doing a job hazard analysis [9]:

- What can go wrong?

doi:10.6342/NTU202401595

- What are the consequences?
- How could it happen?
- What are the contributing factors?
- How likely is it that the hazard will occur?

Based on the Job Hazard Analysis guidelines provided by OSHA, it can be concluded that there are many things that need attention during the job hazard analysis process. However, the main importance can be summarized in what kind of hazard, what are the factors, and how likely a hazard can happen. Therefore, this research aims to integrate historical past accidents recorded by OSHA to a project's schedule in order to get the hazards, sources, and frequencies so that the stakeholders will be able to understand the circumstances of the hazard and the characteristics of the job.

1.3 Structure of Thesis

The thesis is organized as follows. In Chapter 2, we review the literature relevant to our work, namely job hazard analysis, hazard identification, as well as the use of natural language processing (NLP) in construction safety. We additionally share our insights gained from the literature review. In Chapter 3, we introduce the main aspects of NLP driven safety scheduling including, the inputs, processes, and outputs. The gist of our work is presented in Chapter 4. In it, we explain the rationale behind our approach, the selected method for implementing it, as well as the architecture of our proposed model. Chapter 5 focuses on applications of the model through demonstration. First is to integrate the schedule to the Natural Language Processing system, determine the types of hazards that could happen in an activity, and then determine the source for the hazards,

doi:10.6342/NTU202401595

after that calculation for frequency of hazards and frequency of source of hazards. Finally, in Chapter 6, we draw conclusions based on our results and envision further directions for future works.

doi:10.6342/NTU202401595

Chapter 2 Literature Review

2.1 State of Art

Over the years, much research regarding the construction industry has been done, especially research regarding construction safety. The state of art for current technology applications in construction safety can be seen in Table 2.1.

Table 2.1 State of Art for Technology Applications in Construction Safety

Title	**Method**				**Project Phase**		**Description**
	Rule Based Programming	**Computer Vision**	**NLP**	**BIM**	**Planning**	**Monitoring**	
Web-Based Job Hazard Assessment for Improved Safety Knowledge Management in Construction (2019)	V				V		Using a cloud based mySQL database to store all information regarding Job Hazard Analysis and integrate them with schedules so that it is easier for safety personnel to identify and evaluate the potential hazards
Generating Construction Safety Observations via CLIP-Based Image-Language Embedding (2022)		V	V			V	Integrating NLP with computer vision in order to do a safety monitoring based on the condition of the construction site

doi:10.6342/NTU202401595

Table 2.1 State of Art for Technology Applications in Construction Safety (Continue)

Title	Method				Project Phase		Description
	Rule Based Programming	Computer Vision	NLP	BIM	Planning	Monitoring	
Harnessing BIM with risk assessment for generating automated safety schedule and developing application for safety training (2023)	V				V	V	Identifying and assessing hazards in construction activities with FMEA approach and create a safety schedule with visual scripting in Dynamo
Building Information Modeling (BIM) and Safety: Automatic Safety Checking of Construction Models and Schedules (2013)	V			V	V	V	Applying rule-based code checking according to OSHA's guideline to the available BIM model in order to determine which object in which project phase has potential hazards
Application of 4D visualization technology safety management in metro construction (2013)	V	V		V	V	V	Using a rule based programming to integrate some risk assessment into 4D model and update it through monitoring with computer vision
BIM-based fall hazard identification and prevention in construction safety planning (2015)	V			V	V	V	Using BIM and rule based checking to check if in each phase of the construction there is a potential falling hazard

As seen in Table 2.1, a lot of efforts in increasing safety measurements in construction industry, both during the construction phase through monitoring and during

doi:10.6342/NTU202401595

the planning phase through safety planning. One of the methods used to increase safety measurement is through the usage of Natural Language Processing, which allows the contractor to identify hazards through words processing.

2.2 Safety Planning

Safety planning is an important part of designing a project. “Decisions taken at the beginning of the construction process will have a major impact on construction site safety” [7]. Safety planning can be done through safety scheduling and job hazard analysis.

2.2.1 Job Hazard Analysis

In the effort to increase safety in construction, one of the journals in 36th International Symposium introduced a web-based system to analyze potential hazards through Job Hazard Analysis. “By increasing and facilitating access to information, the proposed system can enhance the consistency of JHAs generated throughout the organization, while also ensuring that potential safety risks are not overlooked by less experienced or otherwise biased personnel [10].” The JHA system can be seen in Figure 2.1.

doi:10.6342/NTU202401595

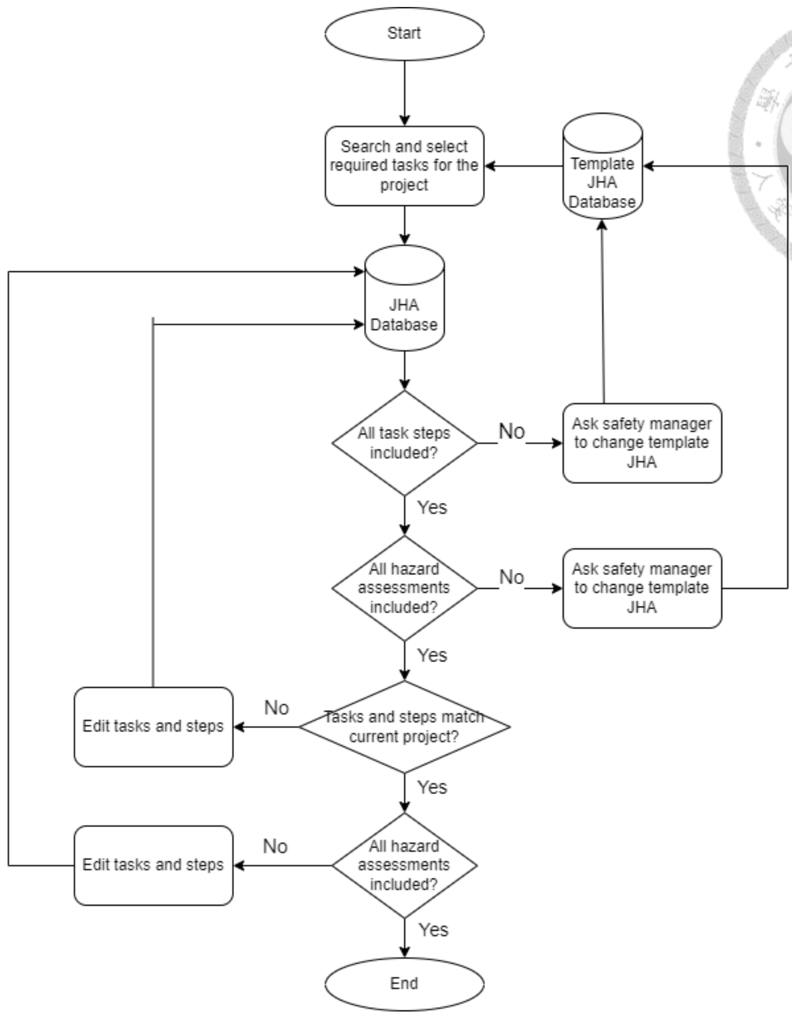

Figure 2.1 Job Hazard Analysis [10]

Through the application of this web-based system, the end user can analyze potential hazards through the type of activities that they listed. The current research enables the safety personnel to search through the job hazard analysis template database for risk assessment and the severity level. The system may still be improved using an automation process using the help of AI or other automation processes such as natural language processing. With the help of automation, the process of job hazard analysis can be shortened and the chance of overlooked potential hazards can be minimized.

doi:10.6342/NTU202401595

2.2.2 Safety Scheduling

By paying attention closely to the schedule made, contractors are able to identify potential hazards that can possibly happen. Potential hazards assessment examples can be seen in Table 2.2.

Table 2.2 Potential Hazards Assessment [11]

Activity Code	Activity	Hazard Code	Hazard	Risk Rank	Risk Level
A1	Excavation / Earthwork	H1	Hit by object	9	Low
		H2	Fall from height to Lower Level	2	High
		H3	Slips & Trips - fall on Same Level	6	Medium
		H4	Caught in or compressed	7	Medium
		H5	Electrocution / Electric shock	1	High
		H6	Transportation accidents	4	High
		H7	Exposure to harmful substances	5	Medium
		H8	Repetitive motion	3	High
		H9	Overexertion / manual handling	10	Low
		H10	Noise, Fire or Others	8	Low

As seen on Table 2.2, some potential hazards can be identified in the early stages of construction. "The safety schedule is designed to explore hazards associated with the activities to be executed on a particular day, as per the project schedule" [11]. There are multiple ways to identify hazards based on the activities in the schedule.

The first option is to do a research study based on a survey done with construction experts. One of the methods used for this is Failure Mode and Effect Analysis (FMEA) which is used to determine the potential hazards from activity and the risk level. The other option is to do research based on the OSHA's reports. Based on the OSHA's database the cause of accidents can be ruled out because some cases are a repetition with fall, struct-by, electrocution, and caught in-between as the top four cause of accidents.

doi:10.6342/NTU202401595

2.3 Natural Language Processing

2.3.1 Natural Language Processing Introduction

Natural Language Processing (NLP) is one of the many applications from artificial intelligence that enables retrieval of information of human language by computer. The functionality of NLP also ranges from analyzing speeches, translating (deep translation), summarizing, recognizing speech, problem solving, etc. [12]. Because of the capability that NLP provides, NLP offers many practical usages in this modern era, ranging from voice assistant application in mobile phones and computers, generative chatbot, search engines, translation tools, and even grammar checking. Figure 2.2 shows how NLP works.

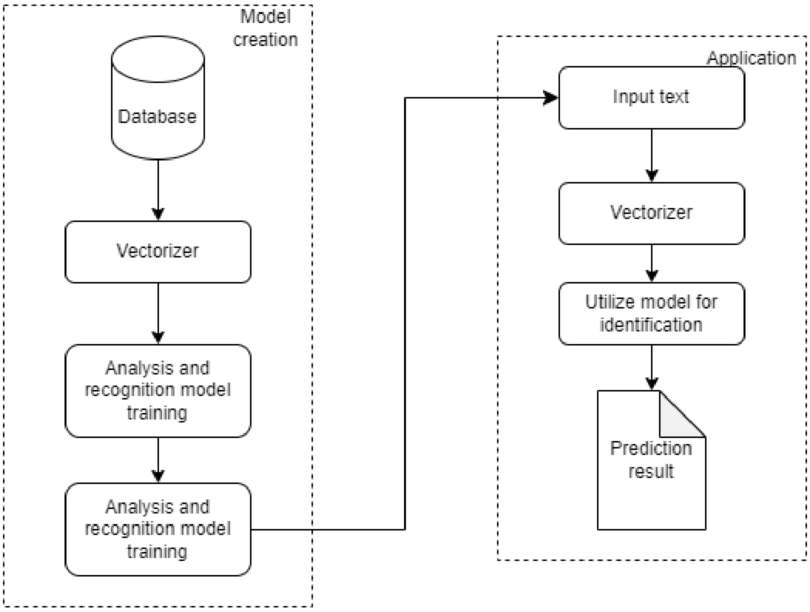

Figure 2.2 Natural Language Processing Model and Application [13]

As seen in Figure 2.2 the NLP utilizes vectorizer to change wording to identifiable numbers and algorithms that can later be used as a base for prediction results. Even though the base formula for NLP usage may be the same, the way to utilize NLP is different depending on the goal and the type of NLP. There are multiple phases in NLP, from

doi:10.6342/NTU202401595

lexical analysis (for words and phrases), syntactic analysis (processing words and identifying relationships), semantic analysis (creating a description), consolidation of speeches, and pragmatic analysis [12].

NLP has undergone a lot of development since it first got released. The development history for NLP can be seen in Figure 2.3.

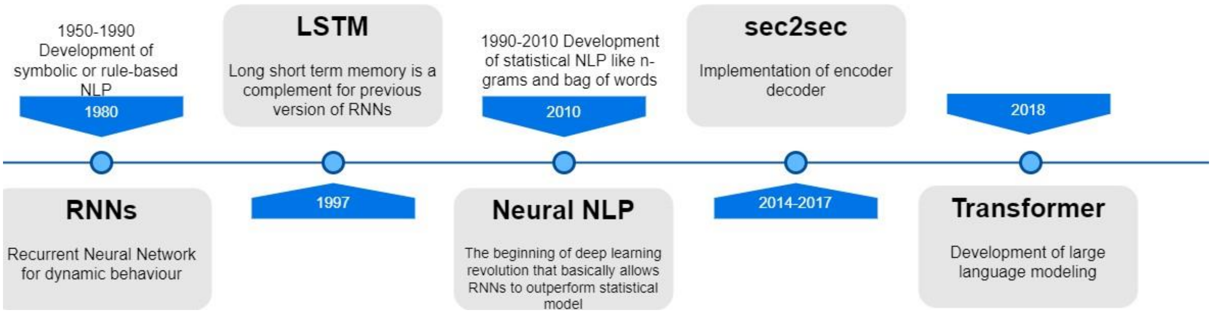

Figure 2.3 NLP Development Timeline [14]

As seen in Figure 2.3 NLP itself consisted of many models from Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Neural NLP, Term Frequency-Inverse Document Frequency (TF-IDF), pretrained model (classification model), and Large Language Modeling (LLM). The usage of these different types of NLP also varies depending on the type of dataset and the goal of the NLP model itself, most of the time newer models don't necessarily mean better quality, every NLP model has each own strength and weakness.

2.3.2 Natural Language Processing Application in Construction Safety

As the economy grows, the construction industry also grows with it, which means a number of constructions are going to happen in the future. OSHA has a database of severe injury reports which consists of texts mentioning the cause of accidents and other information related to the accident. Figure 2.4 shows one of many NLP usages in construction safety.

doi:10.6342/NTU202401595

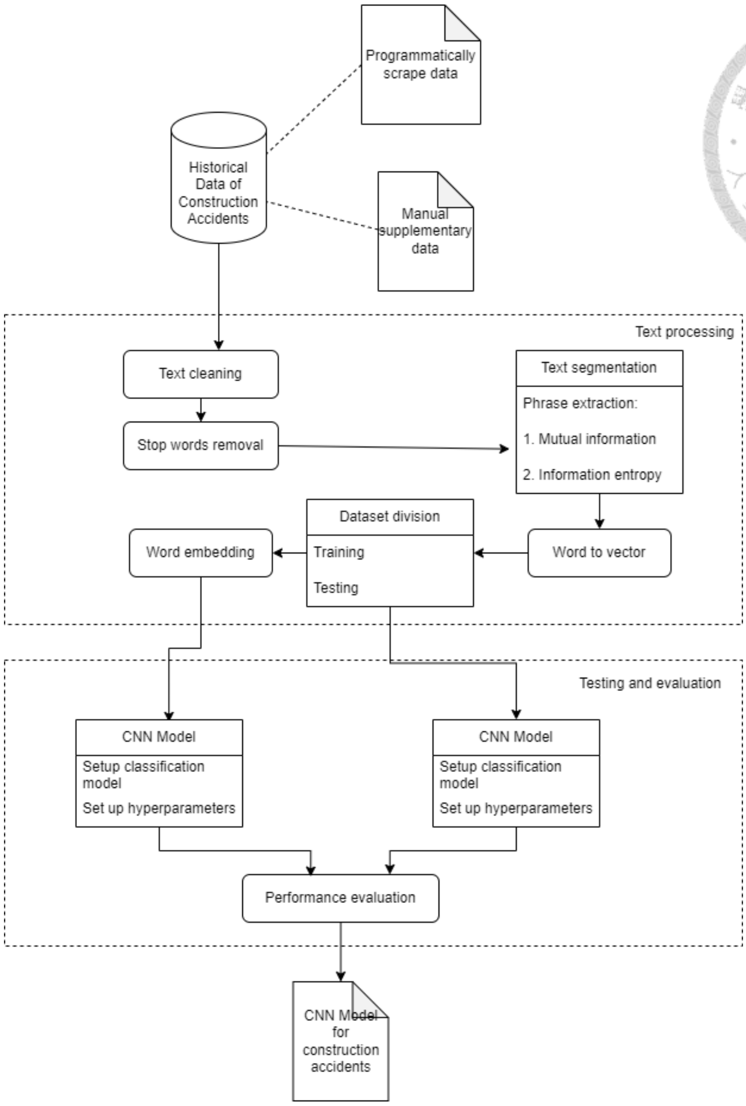

Figure 2.4 How Natural Language Processing Works [15]

One of the reasons NLP is needed is because sometimes manually analyzing unstructured or semi structured textual data can be an exhausting task to do [15]. The automation process will enable faster and more accurate text processing and knowledge acquisition. One of the NLP methods that can be used is convolutional neural network which specializes in recognition and classification. By changing the words model to a vector model and creating a data set consisting of training and testing set, a classification of potential hazards based on text can be created.

doi:10.6342/NTU202401595

Another example of NLP utilization in construction safety is using a word tokenization to make an iteration or prediction based on accident narratives. The process can be seen in Figure 2.5.

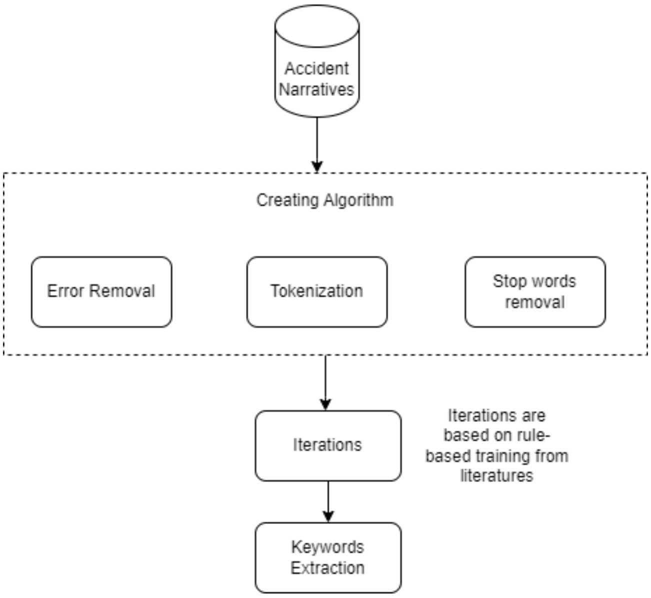

Figure 2.5 NLP Usage in Hazard Classification [16]

As shown in Figure 2.5, NLP can also be used for keywords extraction in order to categorize and identify the most common hazards based on the repetition of keywords in accident narratives. The top 3 results obtained from this research are truck with 1738 words frequency, roof with 907 words frequency, and machine with 551 words frequency which means using the NLP user can be warry of what kind of accidents that usually happen in the construction site [16].

doi:10.6342/NTU202401595

Chapter 3 Methodology

To be able to create an automation of hazard analysis using the construction schedule as a base of identification, the Natural Language Processing is needed. The proposed method will be used to create the automation of hazard identification in this paper. In summary, the process will be divided into several parts, starting from inputting the schedule, extracting important keywords, applying the NLP process, calculating similarity score, filtering the output, calculating the frequency of accidents and sources. This research's methodology can be seen in Figure 3.1.

doi:10.6342/NTU202401595

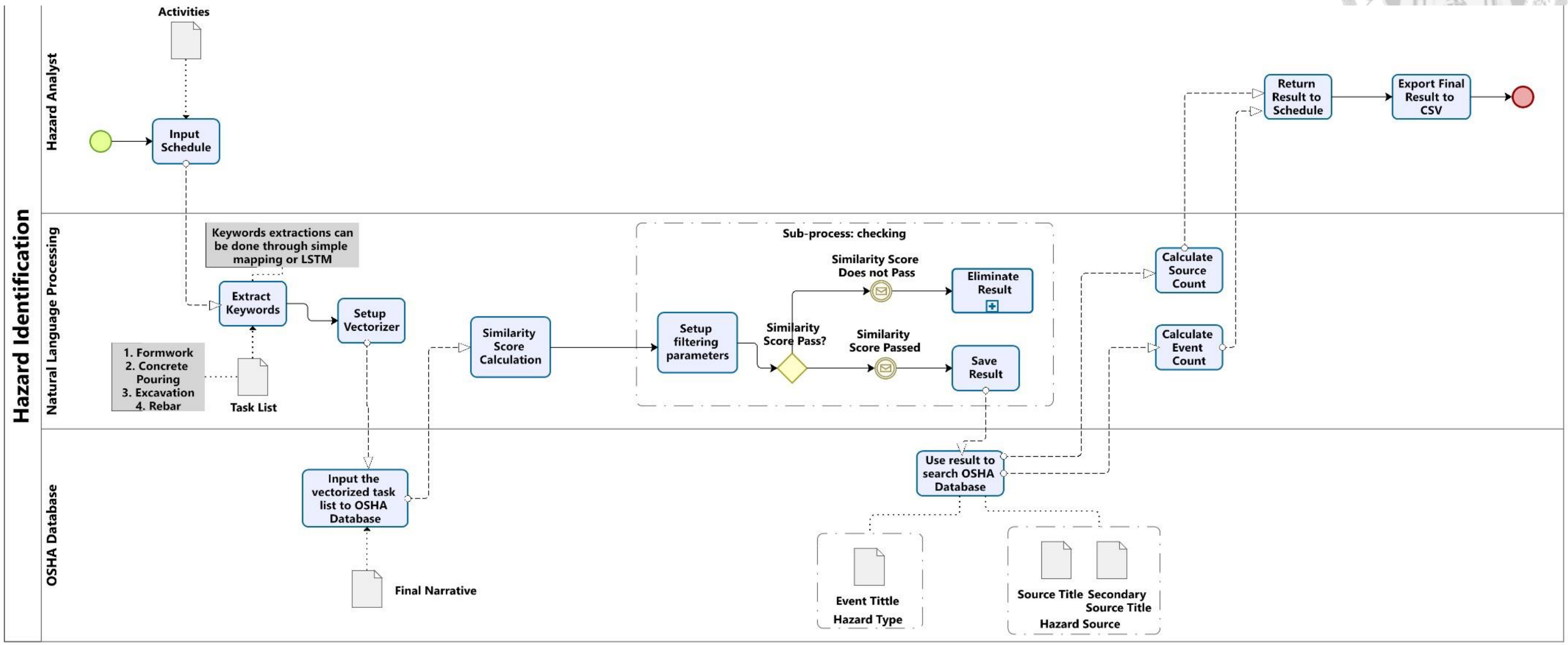

Figure 3.1 Proposed Method

doi:10.6342/NTU202401595

3.1 Natural Language Processing Application

In order for the program to be able to understand the words inside the schedule, keywords extraction must be done. There are several ways to do keywords extraction: keywords mapping and deep learning. For keywords mapping, if the number of important keywords is not big enough, manual keywords mapping using csv or excel can be done. However, if the number of keywords needed are big, instead of the usual keywords mapping, mapping can be assisted by using the Long-Short Term Memory (LSTM). This other method can be done by training the system to be able to locate which keywords can be used in the next phase (NLP application phase).

In this research the Natural Language Processing method chosen as the main approach is Term Frequency – Inverse Document Frequency (TF-IDF). Although TF-IDF can be seen as a more traditional approach in NLP, when developing a concept such as the one being done in this research, TF-IDF is more suitable compared to a more modern approach. This is because of the availability of datasets, and datasets from OSHA are compiled in a simpler format, in which if a more modern approach were to be used, it would not work because in order for it to work, features and labels are needed. The comparison between TF-IDF and other methods can also be seen in Table 3.1.

Table 3.1 Accuracy Comparison Between NLP Models [17]

Method	Accuracy
Proposed 3-Tier CNN Model	98.50%
TF-IDF	98.08%
LSTM With Custom Embedding	97.78%
LSTM	97%
CNN With Char Level Embedding	94%
Basic CNN	92.70%

doi:10.6342/NTU202401595

Table 3.1 shows difference in accuracy between NLP models when used for identifying toxic comments [17]. Although the usage of NLP itself is not the same as what this research is doing, the core concept of classification is the same which in this research it classifies hazard types based on activities from schedule as an input and searches through the Final Narrative. From this table it can also be concluded that even if TF-IDF is considered as a traditional model for NLP, the accuracy remains one of the highest because of the similarity-based calculation.

The TF-IDF is divided into 2 separate processes: TF and IDF. The TF formula is as shown in below:

$$\text{TF(word,doc)} = \frac{\text{Frequency of word} \in \text{the doc}}{\text{No. of words} \in \text{the doc}} \quad (1)$$

where the number of times a word shows up is divided by the number of words inside a document. This formula can be interpreted as: the more often a word shows up in a document, the more important the word is. Meanwhile, the IDF formula is as follows:

$$\text{IDF(word)} = \log_e(1+\frac{\text{No. of docs}}{\text{No. of docs with word}}) \quad (2)$$

where the number of documents is divided by the number of documents that contains the keyword and added by one in a logarithmic calculation. Different from the first formula, Formula 2 is calculating the inverse document frequency, which can be interpreted as: calculating the importance of a word inside the corpus or documents, which means if more documents show the same word, the importance of that word becomes less significant. The implementation of TF-IDF can be seen in Table 3.2.

Table 3.2 TF-IDF Implementation Example [18]

Docs/Words	the	movie	of	pair	was	a	wont	mind
D1	1	1	0	0	1	1	0	0
D2	0	1	1	0	0	0	0	0
D3	0	0	1	1	0	0	0	0
D4	0	0	0	1	0	0	1	1

doi:10.6342/NTU202401595

As seen in Table 3.2, some words show up in more than one document, which means the importance of that word also decreases. The purpose of this is to eliminate words that are often used like prepositions, pronouns, and other repetitive words. To sum it up, "TFIDF is a kind of Natural Language Processing that quantifies the importance of a word within a document relative to its occurrence across a collection of documents, aiming to highlight terms that are both frequent in a specific document and unique to it in comparison to the broader corpus" [18].

3.2 Similarity Score Calculation and Filtering

After the TF-IDF is applied, it is important to check the similarity score to be able to validate whether the result of the Natural Language Processing is to the liking. To calculate the similarity score, first the cosine similarity must be calculated using the following formula:

$$sim(x, y) = \frac{x \cdot y}{\|x\| \|y\|} \quad (3)$$

where $\|x\|$ is the Euclidean norm of vector $x = (x_1, x_2, ..., x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + ... + x_p^2}$ which is known as the length of the vector, the same also works with $\|y\|$ [19]. Once the similarity score has been calculated, the user will be able to check which result is suitable for the research and which one is not. By doing the steps mentioned above, results can be filtered, and more precise answers can be obtained.

In order to obtain a precise result, there are many factors that need to be considered. The factors are the scope of data and the accuracy or precision of the data. In this research, the goal is to create a hazard identification including the frequency and source of hazard.

doi:10.6342/NTU202401595

So aside from accuracy, the scope of data will also need to be considered. The method chosen for this calculation will be F1, precision, recall, and average precision. The formula of precision is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

where the number of true positives (TP) is divided by the total of true positives added by false positives (FP). The positive word indicates all the results above the threshold while the true and false indicates whether the result is relevant or irrelevant. Precision is a formula that is used to display accuracy, precisely the accuracy of the data above threshold. The higher the precision means that it is more likely for the data that is shown or passed the threshold to be correct. Usually, the bigger the data size the lesser the precision will become. This happens because the more data available, the more likely NLP will make a mistake. The formula of recall is as seen below:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

where the number of true positives is divided by the total of true positives and false negatives. False negatives indicate all documents that are supposed to be relevant but are under the picked threshold. Recall is a formula that is used to determine the scope or the size of data above the threshold. As seen in the formula, it is calculated by counting the number of true positives compared to all the data that is supposed to be correct. The bigger the recall means that the number of data that is detected as correct above the threshold increases. Generally, recall and precision go in a different direction, as the recall goes up, the precision will go down because the larger the scope the more likely for the NLP to make mistakes. For this research, recall should be prioritized, considering that this

doi:10.6342/NTU202401595

research aims to create a hazard identification list, which means that the more types of hazards identified the better it is. The F1 formula is as follows:

$$F1 = \frac{2\times(\text{Precision}\times\text{Recall})}{(\text{Precision}+\text{Recall})} \tag{6}$$

where two times precision times recall is divided by the total of precision and recall. While recall may be the most important element to measure this research's goal, F1 also plays an important part to balance out the precision and recall. Even if the main goal is to identify as many hazards as possible, it is also important to maintain the quality of the data obtained, which is why both F1 and precision cannot be overlooked in this research. By considering all three elements, a similarity score threshold for filtering the result can be picked out and the result will have both a good scope and an accurate prediction.

The other formula that is needed is average precision which can be seen in the formula below:

$$\text{Average Precision} = \Sigma_n (R_n - R_{n-1}) P_n \tag{7}$$

where the recall score (R_n) is deducted by previous recall score in previous threshold (R_{n-1}) and then timed by the previous precision score (P_n) and summed to get the average precision score. The average precision score is important to determine whether a language model is robust or not, which means if the average precision is high then the model is usable because an update on the dataset will not influence the model a lot.

doi:10.6342/NTU202401595

Chapter 4 Results

4.1 OSHA Database

As mentioned in chapter 2, one of the steps to perform Job Hazard analysis is to review past accidents. For this purpose, the OSHA database of severe injury report is the most suitable database. The dataset itself has a comma separated value (CSV) format which contains more than 80,000 rows of accident information. However, the dataset itself contains all kinds of industry, not just civil engineering and construction industries, but also farming and other industries. Which is why the data is then sorted out based on the type of industry by sorting out the company's name.

The data was taken from the OSHA website which recorded accident reports, mainly severe injury reports from January 2015 to February 2023 period. Even though the data is called severe injury reports, the report itself was compiled from multiple resources. According to the OSHA website, the data was collected from:

- Report of fatality or severe injury
- Regulatory text – reporting fatalities, hospitalizations, amputations, and losses of an eye as a result of work-related incidents to OSHA
- Year one of OSHA's severe injury reporting program: an impact evaluation
- Worker fatalities reported to federal OSHA

doi:10.6342/NTU202401595

- Fatality reports (archived)

The OSHA database contained multiple columns that includes information on many levels, from detailed to summarized information. The contents can be described in Table 4.1.

Table 4.1 OSHA Database Content

Column Title	**Description**
ID	ID of incident
UPA	Unprogrammed activity ID
Event Date	Date of accident
Employer	Related company
Address 1	Main address
Address 2	Second address
City	City of accident
State	State of accident
Zip	Zip code
Latitude	Latitude coordinate
Longitude	Longitude coordinate
Primary NAICS	Code of industry
Hospitalized	Number of hospitalized workers
Amputation	Number of amputated workers
Inspection	Inspection ID
Final Narrative	Description of accident
Nature	ID of injury's nature
Nature Title	Types of injury (ex: fractures, etc.)
Part of Body	ID of body part
Part of Body Title	Injured body part description
Event	ID of accident's category
Event Title	Accident's categorization
Source	ID of accident's source
Source Title	Primary source of accident
Secondary Source	Secondary source of accident

doi:10.6342/NTU202401595

Based on the OSHA database, multiple columns of data were chosen and used as a database for the Natural Language Processing system. Table 4.2 showcases the datasets used in this research's NLP system.

Table 4.2 Datasets Used in the NLP System

Dataset Needed	**Data Example**
Final Narrative	"On September 28, 2015, at approximately 9:10 a.m., a crane was being used to move a rebar mat into position when one of the lifting straps failed, causing the mat to fall. An employee was beneath the mat as it fell; he became pinned, sustaining broken legs, a punctured lung, and facial lacerations."
Event Title	"Struck by object falling from vehicle or machinery- other than vehicle part"
Source Title	"Structural metal materials, n.e.c."
Secondary Source Title	"Cranes, unspecified"

4.2 Schedule Input and Keywords Extraction

The schedule used for this research is a simple schedule about a building renovation that consists of multiple repetitive activities. The schedule used can be seen in Table 4.3.

doi:10.6342/NTU202401595

Table 4.3 Input Schedule

Activity ID	Activity Name	Start	Finish
10	Furniture removal	05/06/2023 8:00	08/06/2023 16:00
100	Demolishing works for redundant elements	09/06/2023 8:00	15/06/2023 16:00
1000	1F-Laying out walls	16/06/2023 08:00	17/06/2023 10:00
1010	1F-Binding formworks for walls	17/06/2023 10:00	18/06/2023 12:00
1020	1F-Installing rebar for walls	18/06/2023 12:00	22/06/2023 10:00
1030	1F-Casting walls	22/06/2023 10:00	23/06/2023 12:00
1040	1F-Removing formwork for walls	23/06/2023 12:00	24/06/2023 14:00
1100	1F-Wall finishing	24/06/2023 14:00	28/06/2023 12:00
1200	1F-Floor tiling	28/06/2023 12:00	02/07/2023 10:00
1300	1F-Lighting installment	28/06/2023 12:00	30/06/2023 16:00
1400	1F-Ceiling installment	01/07/2023 8:00	03/07/2023 12:00
1500	1F-Window installment	03/07/2023 12:00	04/07/2023 14:00
1600	1F-Door installment	03/07/2023 12:00	05/07/2023 16:00
1700	1F-Green island excavation	28/06/2023 12:00	30/06/2023 16:00
1750	1F-Green island backfilling	01/07/2023 8:00	03/07/2023 12:00
1800	1F-Railing installment	28/06/2023 13:00	03/07/2023 13:00
2000	2F-Laying out walls	28/06/2023 12:00	29/06/2023 14:00
2010	2F-Binding formwoks for walls	29/06/2023 14:00	30/06/2023 16:00
2020	2F-Installing rebar for walls	01/07/2023 8:00	04/07/2023 14:00
2030	2F-Casting walls	04/07/2023 14:00	05/07/2023 16:00
2040	2F-Removing formwork for walls	06/07/2023 8:00	07/07/2023 10:00
2100	2F-Wall finishing	07/07/2023 10:00	10/07/2023 16:00
2200	2F-Floor tiling	11/07/2023 8:00	14/07/2023 14:00
2300	2F-Lighting installment	11/07/2023 8:00	14/07/2023 14:00
2400	2F-Ceiling installment	14/07/2023 14:00	17/07/2023 10:00
2500	2F-Window installment	17/07/2023 10:00	18/07/2023 12:00
2600	2F-Door installment	17/07/2023 10:00	19/07/2023 14:00
2700	2F-Railing installment	19/07/2023 14:00	20/07/2023 16:00
3000	3F-Laying out walls	21/07/2023 08:00	22/07/2023 10:00
3010	3F-Binding formwoks for walls	22/07/2023 10:00	23/07/2023 12:00
3020	3F-Installing rebar for walls	23/07/2023 12:00	27/07/2023 10:00
3030	3F-Casting walls	27/07/2023 10:00	28/07/2023 12:00
3040	3F-Removing formwork for walls	28/07/2023 12:00	29/07/2023 14:00
3100	3F-Wall finishing	29/07/2023 14:00	02/08/2023 14:00

doi:10.6342/NTU202401595

3200	3F-Floor tiling	02/08/2023 14:00	06/08/2023 12:00
3300	3F-Lighting installment	02/08/2023 14:00	07/08/2023 14:00
3400	3F-Ceiling installment	07/08/2023 14:00	10/08/2023 10:00
3500	3F-Window installment	10/08/2023 10:00	11/08/2023 12:00
3600	3F-Door installment	10/08/2023 10:00	12/08/2023 14:00
3700	3F-Railing installment	12/08/2023 14:00	13/08/2023 16:00
4000	RF-Floor tiling	14/08/2023 08:00	16/08/2023 12:00
4100	RF-Ceiling installment	14/08/2023 08:00	16/08/2023 12:00
4200	RF-PV panels installment	22/08/2023 12:00	31/08/2023 10:00
4300	RF-Door installment	31/08/2023 10:00	02/09/2023 16:00
4400	RF-Railing installment	31/08/2023 10:00	03/09/2023 16:00
4500	RF-Roof canopy installment	16/08/2023 13:00	21/08/2023 13:00
5000	Plants wall installment	04/09/2023 8:00	07/09/2023 14:00
5100	Exterior wall painting	07/09/2023 14:00	11/09/2023 12:00
5200	Wood facade installment	11/09/2023 12:00	15/09/2023 09:00
5300	Site cleaning and decoration	15/09/2023 09:00	23/09/2023 15:00
5400	Funiture installment	23/09/2023 15:00	26/09/2023 11:00

Before the schedule is then used as an input for the NLP system, the keywords must first be extracted. There are multiple ways to extract important keywords from the schedule, the first one is by using simple keywords mapping. Using this method means the user must create a list of activities or keywords that can be taken off from the schedule to be used as an input for the TF-IDF search base. Figure 4.1 showcases keywords used in the mapping process.

	Keywords Schedule	Keywords OSHA
0	Formwork, Form, Formworks, Forms	Formwork, Form, Formworks, Forms
1	Pouring Concrete, Concrete Casting, Casting, C...	Pouring Concrete, Pouring, Concrete Pump, Pump...
2	Excavation, Digging	Excavation, Digging
3	Rebar, Reinforcement, Rebars	Rebar, Reinforcement, Rebars, Reinforcements

Figure 4.1 Keywords Mapping

doi:10.6342/NTU202401595

As seen in Figure 4.1, before processing the keywords extracted from the schedule, the keywords must first be converted into a format that is suitable for the TF-IDF system. The reason is because there are certain rules that must be followed when creating an accident report and uploading them to OSHA website, which is why although the case may be different from one to another, the writing style or format will still have to follow the OSHA guidelines according to the codes that have been prepared by OSHA beforehand. In order to do this process, there are multiple steps that must be taken (see Figure 4.2)

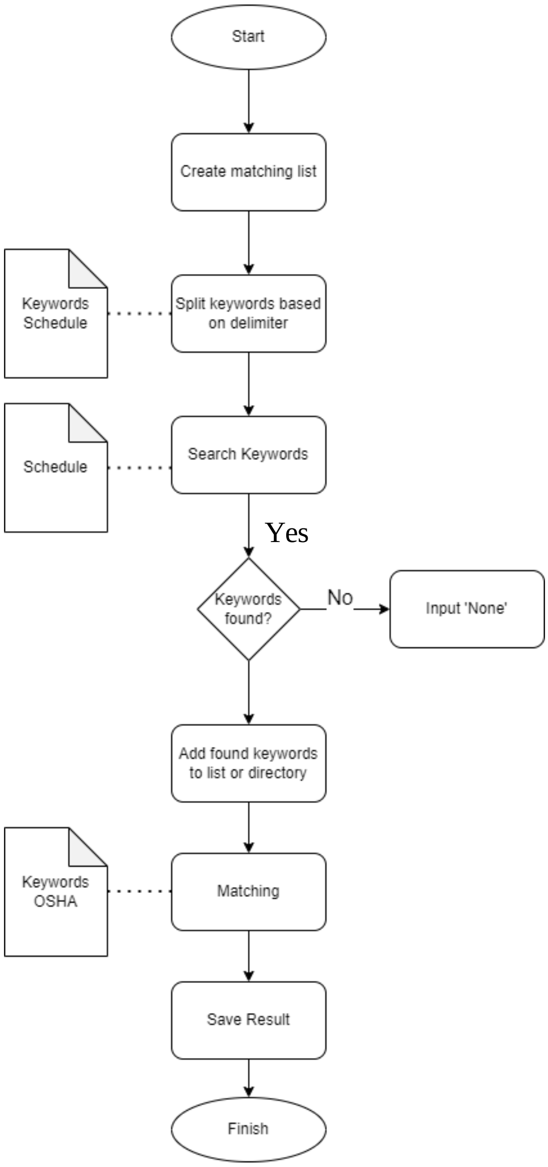

Figure 4.2 Keywords Mapping Procedure

doi:10.6342/NTU202401595

Once the mapping process is finished, the result can then be used as an input for the main process, which is the TF-IDF system.

The second option for this mapping process is by using another language modeling, like Long Short-Term Memory (LSTM). This option can be used for a much more complex schedule. The purpose of this method is to obtain a fully automated labeling process through training and testing process. Basically, this method allows the mapping process to be fully automated using different kinds of classification approaches. However, the reason this method was not chosen for this thesis is because in this research the focus point is more on creating a prototype, which only consists of 4 activities (formwork, rebar, concrete pouring, and excavation).

The third option may be slightly different from the other two. Using the generative artificial intelligence (GPT AI), some important information can be extracted from both the OSHA Database and schedule directly and using python, the result can be used to match compatible information. Keywords extraction using GPT AI can be seen in Figure 4.3.

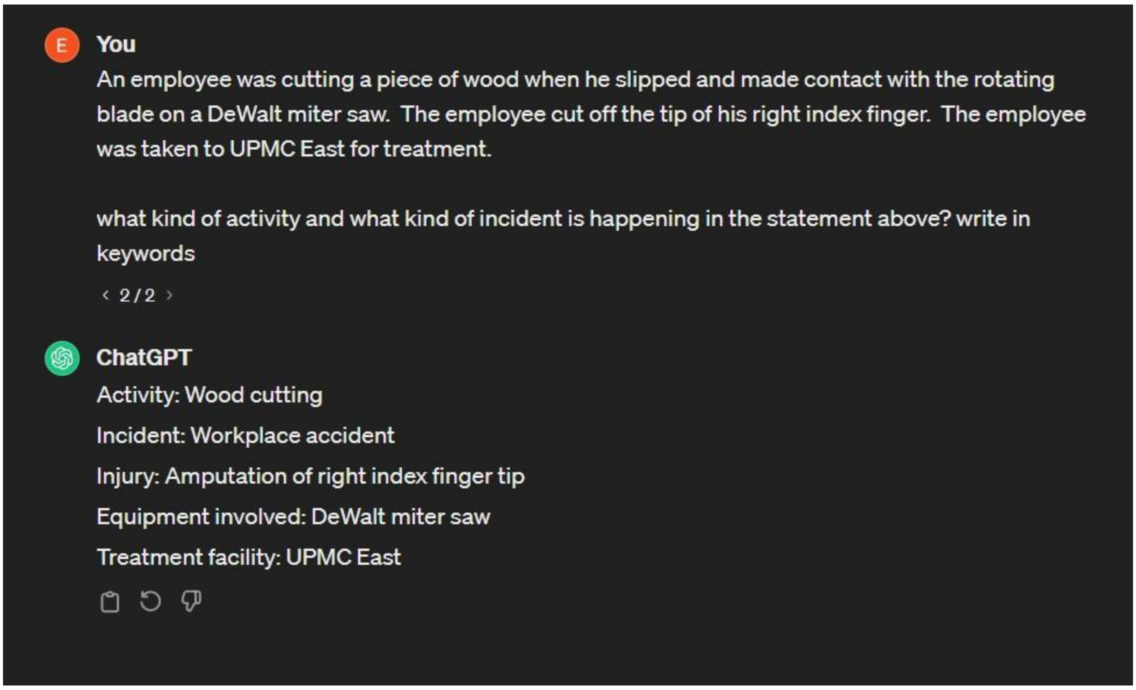

Figure 4.3 Keywords Extraction Using GPT AI

doi:10.6342/NTU202401595

As seen in Figure 4.3, GPT AI is able to extract some key information from the OSHA database that can later on be used to match with the activities from the schedule. However, note that because the GPT AI itself is not perfect yet, there will be a possibility for hallucinations which will give out a fake result when the result prompted is not available.

Out of the 3 methods for keywords mapping, this research focuses more on the manual keywords mapping in order to be able to create a simple yet fulfilling prototype of the hazard identification system. The keywords extraction process is as seen in Figure 4.4.

doi:10.6342/NTU202401595

```
0 None
1 None
2 None
3 {'index': 0, 'keyword': 'Formwork'}
4 {'index': 0, 'keyword': 'Rebar'}
5 {'index': 2, 'keyword': 'Casting'}
6 {'index': 0, 'keyword': 'Formwork'}
7 None
8 None
9 None
10 None
11 None
12 None
13 {'index': 0, 'keyword': 'Excavation'}
14 None
15 None
16 None
17 {'index': 1, 'keyword': 'Form'}
18 {'index': 0, 'keyword': 'Rebar'}
19 {'index': 2, 'keyword': 'Casting'}
20 {'index': 0, 'keyword': 'Formwork'}
21 None
22 None
23 None
24 None
25 None
26 None
27 None
28 None
29 {'index': 1, 'keyword': 'Form'}
30 {'index': 0, 'keyword': 'Rebar'}
31 {'index': 2, 'keyword': 'Casting'}
32 {'index': 0, 'keyword': 'Formwork'}
33 None
34 None
35 None
36 None
37 None
38 None
39 None
40 None
41 None
42 None
43 None
44 None
45 None
46 None
47 None
48 None
49 None
50 None
```

Figure 4.4 Keywords Extraction Process

The keywords extraction result can be seen in Figure 4.4. It can be seen that designated keywords or prepared keywords are found in a number of activities in the schedule. Because the system made is a prototype so activities other than rebar,

doi:10.6342/NTU202401595

formwork, concrete pouring, and excavation will not be added into the index, which is why it is displayed as 'None.' Next step is keywords conversion (see Figure 4.5)

```
Activity Index: 3
  Matched Keyword: Formwork
  Keywords OSHA: Formwork, Form, Formworks, Forms


Activity Index: 4
  Matched Keyword: Rebar
  Keywords OSHA: Rebar, Reinforcement, Rebars, Reinforcements


Activity Index: 5
  Matched Keyword: Casting
  Keywords OSHA: Pouring Concrete, Pouring, Concrete Pump, Pump, Concrete Pour, Pour, Pumping Concrete, Pumping


Activity Index: 6
  Matched Keyword: Formwork
  Keywords OSHA: Formwork, Form, Formworks, Forms
```

Figure 4.5 Keywords Conversion

As seen in Figure 4.5, after the keywords are extracted from the schedule activities, they are converted to a format that is understandable by the OSHA Database (according to OSHA report writing guidelines). After that, the keywords will be saved to an index that will later be used as a search base for the TF-IDF system.

4.3 TF-IDF

4.3.1 Setting Up TF-IDF

The TF-IDF method consisted of several steps, from setting up the vectorizer, printing out results, checking, similarity score filtering, to printing out results, frequency, and source of accidents. The steps mentioned can be seen in Figure 4.6.

doi:10.6342/NTU202401595

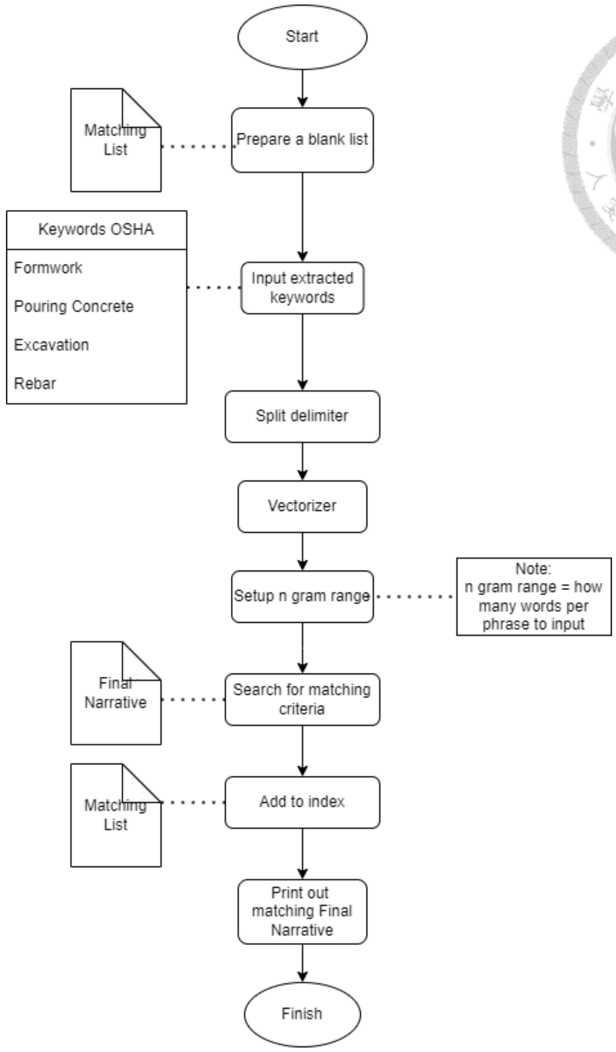

Figure 4.6 TF-IDF Process

As seen in Figure 4.6, the first step is to create a blank list, the blank list will then be used to store matching keywords in Final Narrative. After that, to input the prepared keywords, the keywords are then separated to create multiple keywords by dividing based on the delimiter (,). After that keywords are put into a vectorizer so that the NLP will be able to recognize the keywords by converting them to a number-based code. To get a better result, the NLP must understand not just words, but also phrases, to be able to do that, n-gram needs to be set up. The n-gram set up process can be seen in Figure 4.7.

doi:10.6342/NTU202401595

```
# List of keywords that require a specific ngram_range for TfidfVectorizer
special_keywords = ["pouring concrete", "pouring", "concrete pump", "pump", "concrete pour", "pour", "pumping concrete", "pumping"]

    # Create a TF-IDF vectorizer with ngram_range condition for special keywords
    if any(kw in keywords1 for kw in special_keywords):
        vectorizer = TfidfVectorizer(ngram_range=(1, 2))
```

Figure 4.7 Setting Up N Gram Range

As seen in Figure 4.7, there are two parameters that need to be set in an ngram-range. The first parameter is used to set up the number of phrases, while the second parameter is used to set up the number of words in a phrase. It can also be seen in Figure 4.7 that pouring concrete is differentiate from the other activities because the number of ngram-range needed will also be different. Using ngram-range (1, 2) means that the result will not only be looking for single words but also bigrams, for example using the phrase "concrete pouring" will result in the phrase "concrete pouring" and "concrete", this setting is suitable for this research because it can not only look for related words but also words that are similar or related to "concrete". The reason that some words are written more than once (for example "pouring concrete" and "concrete") in the keywords list is because in order to produce a balanced result, some words need to be repeated. For example, in the special keywords list, the words "pouring", "pour", "pump", and "pumping" are written twice to prevent the word "concrete" from dominating the search because of the 1,2 n-gram setting. The n-gram package used for this procedure is directly adopted from the TF-IDF package provided by python. After these procedures are finished, the next step can be started.

doi:10.6342/NTU202401595

4.3.2 Similarity Score Calculation

In order to get an accurate result, similarity score calculations are needed. The calculations are based on the TF-IDF formula. Using targeted keywords that have been extracted from the schedule, the keywords will be used as a benchmark. To get the calculation score, the formula used is the *cosines* of the TF-IDF formula. The results of the top 10 similarity score calculations can be seen in Table 4.4 to Table 4.7.

As seen from Table 4.4 to Table 4.7, the algorithm that TF-IDF uses searches the keywords based on the similarity and TF-IDF formula. From the highlights it can also be seen that the important keywords are found in the Final Narrative printed by TF-IDF. Similarity calculation works based on how many words are in a sentence and how often the important words are repeated. In most cases, as the similarity score goes down, the accuracy of the prediction will also go down, which is why the filtering process is necessary.

Table 4.4 TF-IDF Formwork

Score	**Final Narrative**
0.3506	"An employee was removing chains from plywood formwork. The formwork was hit by another piece of wood, and the wooden formwork struck the employee on their lower back, resulting in a back contusion."
0.3444	"An employee was preparing to dismantle formwork. With the wedge pins of the formwork removed, the employee attempted to step out onto an outrigger. The employee fell and his harness, tied off to the formwork, brought the form work down on his ankle, breaking the ankle."
0.3325	"An employee was stripping nails from formwork when the formwork shifted and lacerated the employee's left ring finger."

doi:10.6342/NTU202401595

0.2856	"A carpenter employee received multiple fractures to his leg when he was struck by and pinned by a section of concrete form. The employee was climbing down from the top of a stack of concrete forms when he stepped on an L bracket on one of the forms causing the form to move and strike the employee."
0.2780	"An employee was releasing a load of formwork. The form dropped on the employee's thumb, causing the amputation of the thumb tip."
0.2662	"An employee was removing 2-by-4s from bridge forms. A form fell and broke the employee's collarbone and ribs."
0.2573	"An employee was setting a concrete form when the form pinched and amputated their right index finger."
0.2310	"An employee was disassembling concrete wall form panels. The steel form came loose and knocked the employee to the ground. The steel form then fell on his right leg, resulting in a lower right leg fracture."
0.2283	"An employee was climbing down from a formwork wall when he fell approximately 10 feet to the ground and broke his right leg."
0.2193	"An employee was moving forms when his left hand was caught between a shear wall form and an aluminum beam. His left pinky and ring fingers were lacerated."

Table 4.5 TF-IDF Pouring Concrete

Score	**Final Narrative**
0.1925	"Employees were pouring concrete in drill shafts. An overhead power line arced, struck the concrete pump truck, and struck an employee. The employee was shocked and lost consciousness and was hospitalized to treat right hand/fingertip burns."
0.1896	"An employee was pouring concrete when he became disoriented due to heat stress, requiring hospitalization."

doi:10.6342/NTU202401595

0.1819	"An employee was pouring concrete when concrete entered his boots, causing a chemical burn that became infected."
0.1814	"An employee was pouring concrete for a slab and foundation using a boom pump. The hose on the pump broke and fell off the boom, striking the employee and injuring the employee's hip, back, and legs."
0.1809	"An employee was struck on the chest and face by a concrete pump truck hose when the pump operator started the pump."
0.1604	"An employee was working near a concrete pump truck that had a hose and hose clamp connected to the exterior output connection of the truck. The truck was pumping concrete when the hose clamp broke off and struck the employee in the face. The employee suffered facial lacerations."
0.1598	"An employee was standing on a concrete form, pumping concrete. The pump drew in air, and it caused an air hammer in the line, knocking the employee from the form. He suffered broken ribs, a broken arm, and other injuries."
0.1558	"An employee was pouring concrete with a hose. Air got into the line and the hose struck him in the head and neck."
0.1527	"An employee and coworkers were pouring a concrete floor when the floor collapsed about 18 feet. The employee suffered back and rib injuries and was hospitalized."
0.1514	"An employee was preparing a concrete pump to fill a concrete truck hopper when the employee was struck by the concrete truck, resulting in a crushed pelvis."

Table 4.6 TF-IDF Excavation

Score	Final Narrative
0.4720	"A crew was digging with a rubber-tired backhoe to raise a valve can. The excavation was approximately 3.5 feet deep. Dirt from one side of the excavation collapsed into the hole and pinned the injured employee's leg

doi:10.6342/NTU202401595

	against the existing utility in the excavation. The employee was hospitalized for a broken leg."
0.4179	"An employee was working on the edge of a 3-foot excavation. He fell into the excavation, landing on his back and suffering broken vertebrae."
0.3328	"An employee was standing on the bank of an excavation. The side of the excavation gave way and the employee fell in. Concrete pieces fell on top of him, and he sustained a broken leg."
0.3113	"An employee was setting up the excavation with lamp light and fell into the excavation after his glasses fogged up. The employee sustained a fractured rib. "
0.3013	" An employee was repairing a digging chain when he fell backward, fracturing ribs and lacerating his head."
0.2968	"Two employees were digging and dewatering an excavation when one of them came into contact with a live underground power line. One employee suffered an electric shock and left hand burns; the other was also shocked. Both employees were hospitalized. "
0.2808	"An employee was hand digging a hole when an excavator struck his right leg. His right leg had to be amputated at the knee. "
0.2735	"At 1:45 p.m. on June 29, 2020, an employee was installing plumbing at the bottom of an excavation, about 5-7 feet deep, when the side of the excavation collapsed and fell on him. He suffered fractures to the ribs, collarbone, and pelvis and was hospitalized. "
0.2461	"An employee was digging a hole with a shovel when they struck an electric cable and an arc flash occurred, resulting in a burn."
0.2401	"An employee was struck by a light pole during an excavation operation resulting in a back injury."

doi:10.6342/NTU202401595

Table 4.7 TF-IDF Rebar

Score	Final Narrative
0.2670	"The injured employee was pounding rebars into the ground for a concrete form. Another employee bumped into the injured employee causing them to fall forward onto the rebar. The rebar impaled the injured employee on the side of his neck. "
0.1836	"An employee was operating a machine used to bend rebar. The employee was holding a piece of rebar when their fingers were caught between the rebar and the machine's backstop guide resulting in a fingertip amputation. "
0.1797	"An employee was bending and cutting rebar on a rebar bending machine when the machine cycled twice, causing the rebar to bend twice. The rebar punctured the employee's right foot, fracturing the fourth metatarsal. "
0.1527	"An employee was waiting for rebar to be unloaded from a trailer. Straps broke and rebar struck the employee in the head. The employee was hospitalized with a head injury. "
0.1521	"An employee was snipping a piece of wire off of a horizontal, temporary, piece of rebar. The rebar fell and struck the employee's thumb resulting in an amputation. "
0.1466	"An employee was shearing rebar to different lengths. The rebar shear's hold-down clamp crushed the employee's right hand. "
0.1464	"An employee was sandblasting rebar so he could lay new concrete when he tripped on a piece of rebar and sandblasted his right foot."
0.1441	"An employee was descending a 12-foot column form and came down on a section of rebar that was in his path. The piece of rebar impaled his left upper leg. "

doi:10.6342/NTU202401595

0.1428	“An employee was putting rebar in a down cell in a block wall. The employee was electric shocked when the rebar he was working with arced with a power line. “
0.1416	“An employee was about 13' up a cell tower in a spider basket taking steel off of the tower for new reinforcement. The employee fell and sustained lower back fractures. The employee was not tied off at the time. “

4.4 Similarity Score Filtering

Before the filtering process starts there are a few things that need to be done. First is to set up the parameter in order to be able to judge properly whether a data should be in the hazard identification or not, second labeling the data based on the parameters, and finally determining outputs can be calculated (precision, recall, average precision, and F1). For this section, the calculations are done activity by activity in order to ensure that the result produced by TF-IDF is the most suitable with the needs of hazard identification considering different activities have different size of results that need to be filtered. Each activity results are divided into 80% training dataset and 20% testing dataset randomly using phyton. The training mentioned means that the 80% dataset will be used to pick a similarity score threshold while the testing means validation or to validate whether the threshold picked produces acceptable outputs and whether the model is usable for future updates. Parameters, labeling, and calculation process will be shown in Table 4.8 to Table 4.9. Setting up parameters are based on the needs of the schedule, which is why it is important to pay attention to what kind of schedule is used as an input. There are some details that need attention in the schedule: what are the activities, what are the scope of work, what is the level of detail, what are the steps of each activity. Setting up the

doi:10.6342/NTU202401595

parameters can be done by analyzing the information on the input schedule and needs to be done before the similarity score filtering process to ensure the results are objective and accordingly.

Table 4.8 Setting Up Parameters

Activity	Label	Parameters
Formwork	0	Accidents that are not caused by formwork activities or formwork placement (climbing down formwork, rebar accidents, vehicle accidents, etc.)
	1	Accidents directly caused by activities related to formwork (forming a structure, prying open, dismantling, moving, drilling, cutting, stripping, etc.) and accidents caused by improper formwork quality or placement (getting hit, getting knocked out, etc.)
Pouring Concrete	0	Accidents caused by activities outside of concrete pouring process (cutting concrete, precast concrete, removing concrete, cleaning concrete at washing bay, finishing concrete, etc.) and accidents that are unrelated with concrete pouring activities (falling and hitting concrete floor, etc.)
	1	Accidents that are caused by concrete pouring process (setting up pump, spreading concrete, moving wet concrete, etc.)
Excavation	0	Accidents caused by activities outside of excavation (pipe installation, transporting materials and tools from excavation, etc.) and accidents that are caused by other things inside

doi:10.6342/NTU202401595

		the excavation (struck by machine while moving or transporting something, etc.)
	1	Accidents caused by excavation or digging (including placing a sheet on the sides), falling to excavation because of the lack of proper measurements, and accidents caused by bad excavation (landslide, collapsing soil, etc.)
Rebar	0	Accidents caused by activities unrelated to rebar (falling while climbing and getting impaled, formwork, drilling, etc.)
	1	Accidents caused by rebar (cutting, wiring, moving, installing, etc.)

As seen in Table 4.8, the parameters were set up specifically so that the result will be as objective as possible, here are some of the labeling processes based on the similarity score obtained in the previous section. Three 0 labels and three 1 labels will be shown in Table 4.9.

Table 4.9 Parameter Based Labeling

Keywords	**Similarity**	**Final Narrative**	**Label**
formwork, form, formworks, forms	0.234373854	"An employee was climbing down from a formwork wall when he fell approximately 10 feet to the ground and broke his right leg."	0
	0.083836923	"On July 27, 2022, at approximately 2:45 p.m., an iron worker fell approximately 4' 10" to the ground while removing existing rebar in a form wall. The employee lost consciousness and	0

doi:10.6342/NTU202401595

		sustained four broken ribs. The employee was hospitalized."	
	0.026784475	“Two employees were bolting round flanges into a pipeline formation. One employee was torqueing with a ratchet while the other was holding a multiplier that was attached to the ratchet with the use of an extender. There was another 2" flange (which extended further out than the flange that was being bolted) that was in the middle of the flange that they were bolting. They were torqueing the bolts on the flange in sequence/series. When they started the last round of torquing, the employee holding the multiplier got his pinky caught between the multiplier and the center flange. This resulted in a partial amputation of the left pinky finger. The employee had removed skin form the very tip of the finger and the nail bed.”	0
	0.360703637	“An employee was removing chains from plywood formwork. The formwork was hit by another piece of wood, and the wooden formwork struck the employee on their lower back, resulting in a back contusion.”	1
	0.191143669	“An employee was using a pry bar to guide two concrete forms a few centimeters into place when their right middle finger was pinched between the	1

doi:10.6342/NTU202401595

		two forms, resulting in a fingertip amputation."	
	0.128089266	"An employee was helping to form concrete when he became hot and passed out. He was hospitalized for dehydration."	1
pouring concrete, pouring, concrete pump, pump, concrete pour, pour, pumping concrete, pumping (n-gram 1,2)	0.079507527	"Employees were installing a pump at the facility. During the process, the pump shifted and caught one of the employees' fingers, smashing it against the pedestal wall of the pump. Employee was transported to the Memorial Herman Hospital where they determined that the finger had a fracture."	0
	0.070017978	"An employee was part of a four-person crew that was removing a form after a concrete pour was set. A crane tried to pull the form off a concrete column, and the wire rope guideline hung up on something. The employee tried to release the guideline"	0
	0.030973132	"An employee hanging dry wall fell from a 6-foot ladder to the concrete floor. The employee was hospitalized with a spleen injury and rib fracture."	0
	0.194309993	"An employee was pouring concrete when he became disoriented due to heat stress, requiring hospitalization."	1
	0.088128576	"An employee was supervising a concrete pour. The deck that was being poured gave way, and the employee's	1

doi:10.6342/NTU202401595

		right foot and ankle were pinched between two deck boards. The foot and ankle were broken.”	
	0.042383186	“An employee was standing on the top of concrete forms, which were part of the setup for the basement walls of a new house under construction. The employee was directing the flow of concrete into the forms. The hose from the concrete pumper truck moved unexpectedly as a result of a clog in the hose and knocked the employee to the ground in the interior of the basement. The employee suffered minor vertebrae fractures.”	1
excavation, digging	0.294851718	“An employee was repairing a digging chain when he fell backward, fracturing ribs and lacerating his head.”	0
	0.139355765	“On October 10, 2016, at approximately 4:30 p.m., an employee was helping to remove a light plant generator from an excavation. A 9-foot chain was attached to telehandler forks and the light plant was lifted from the excavation and transported to a flatbed pickup truck located below a power line. While standing on the ground, the employee used his hand to guide the light plant onto the 4' high flatbed. The telehandler forks made contact with the overhead 7,200-volt line. The employee received second degree burns to his left hand, abdomen, and leg.”	0

doi:10.6342/NTU202401595

	0.086710373	"On November 6, 2021, an employee was spray painting a petroleum pipeline for rust prevention. The pipeline is supported by wood log cribbing and subject to move when the temperature rises. The pipeline fell and pinned the employee down in the excavation. The employee sustained several broken ribs, breathing obstruction, contusion, and loss of consciousness. He was hospitalized."	0
	0.465101945	"A crew was digging with a rubber-tired backhoe to raise a valve can. The excavation was approximately 3.5 feet deep. Dirt from one side of the excavation collapsed into the hole and pinned the injured employee's leg against the existing utility in the excavation. The employee was hospitalized for a broken leg."	1
	0.290446697	"Two employees were digging and dewatering an excavation when one of them came into contact with a live underground power line. One employee suffered an electric shock and left-hand burns"	1
	0.167544455	"An employee was driving sheet metal into an excavation to prevent cave-in. The metal contacted an electrical line, and the employee was shocked and burned."	1
rebar, reinforcement,	0.079689716	"An employee was walking around the corner of a foundation under	0

doi:10.6342/NTU202401595

rebars, reinforcements		construction, tripped and fell into rebar, suffering a laceration to his inner right thigh.”	
	0.062028055	“An employee was looking for stakes in the ground that marked areas for drilling. He tripped over a stake. When he fell to the ground his left arm hit a piece of rebar resulting in a forearm fracture. The employee was hospitalized.”	0
	0.06032273	“An employee was performing a walk around visual inspection of the crusher recycler when he tripped on a piece of rebar and fell hard on his left hip and forearm. His hip broke and required surgery.”	0
	0.254968498	“The injured employee was pounding rebars into the ground for a concrete form. Another employee bumped into the injured employee causing them to fall forward onto the rebar. The rebar impaled the injured employee on the side of his neck.”	1
	0.123467736	“An employee was in front of a table bender rebar machine, bending rebar. The employee's left middle fingertip was caught in a hole in the rotating top. The fingertip was nicked, resulting in an amputation.”	1
	0.098469916	“An employee developed heat exhaustion while tying rebar in the	1

doi:10.6342/NTU202401595

		afternoon. The employee was hospitalized.”	

The labeling process for the training dataset can be seen in Table 4.9, as mentioned in Table 4.8 the parameters set a clear distinction between the “0” or irrelevant results with the “1” or relevant results. Based on the labeling, the precision, recall, average precision, and F1 are then calculated. The results are as seen in Figure 4.8 to Figure 4.27.

doi:10.6342/NTU202401595

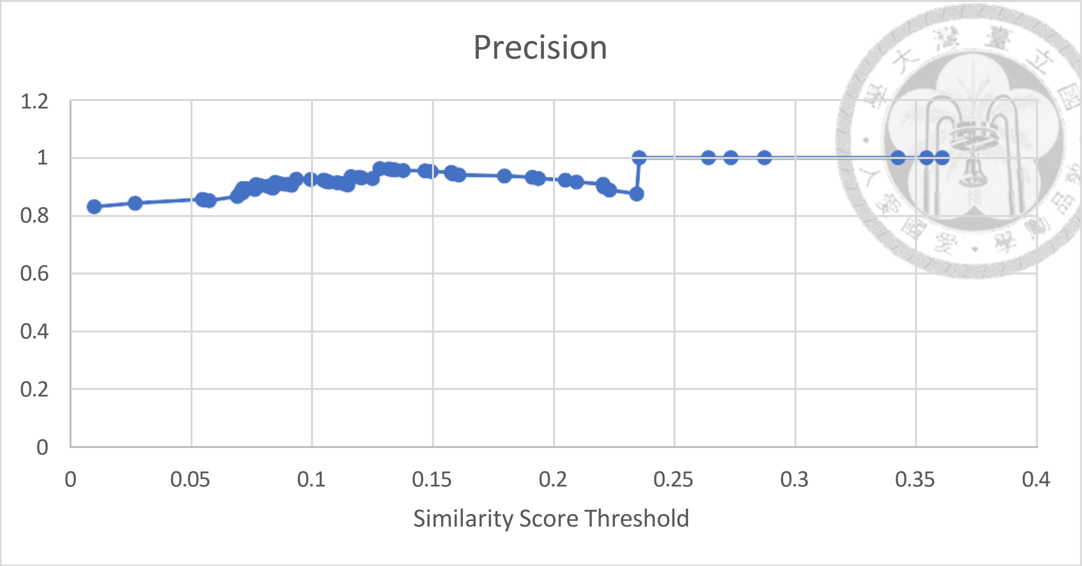

Figure 4.8 Precision for Formwork Training Dataset

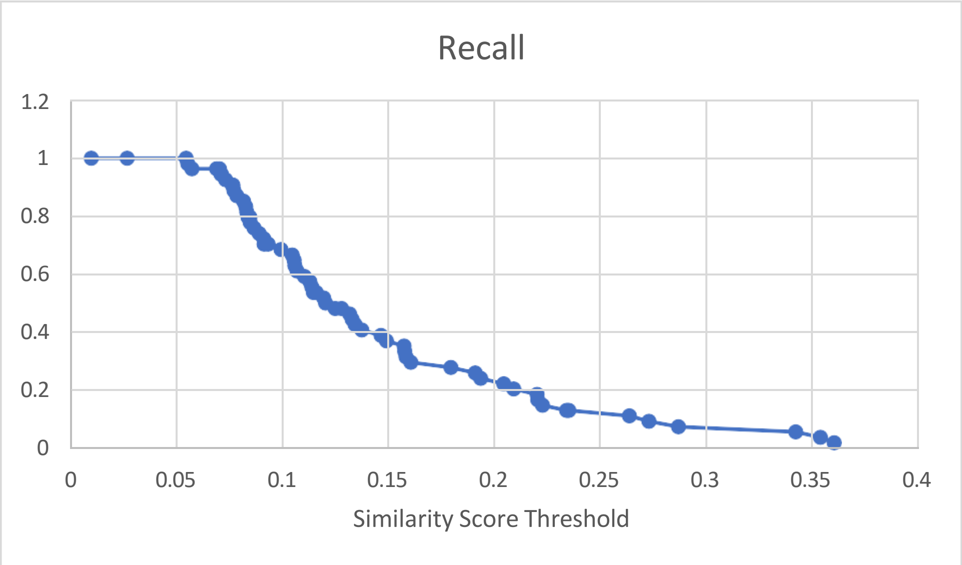

Figure 4.9 Recall for Formwork Training Dataset

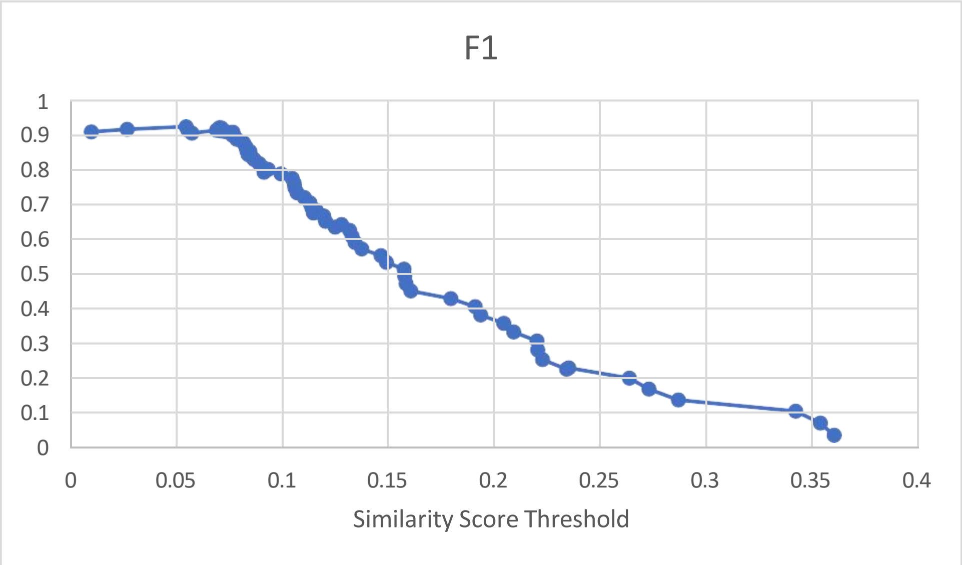

Figure 4.10 F1 for Formwork Training Dataset

doi:10.6342/NTU202401595

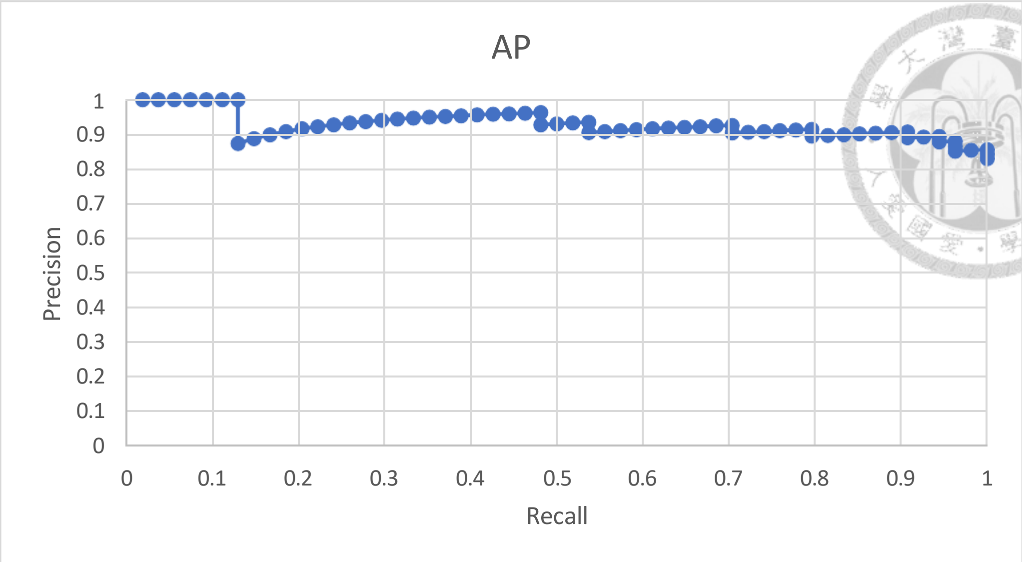

Figure 4.11 Average Precision for Formwork Training Dataset

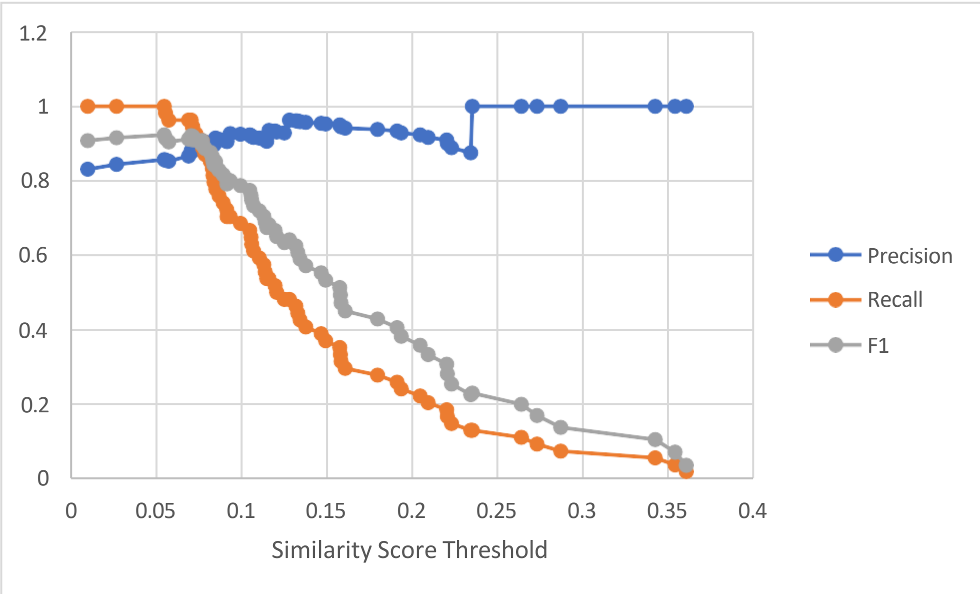

Figure 4.12 Summary of Formwork Training Dataset

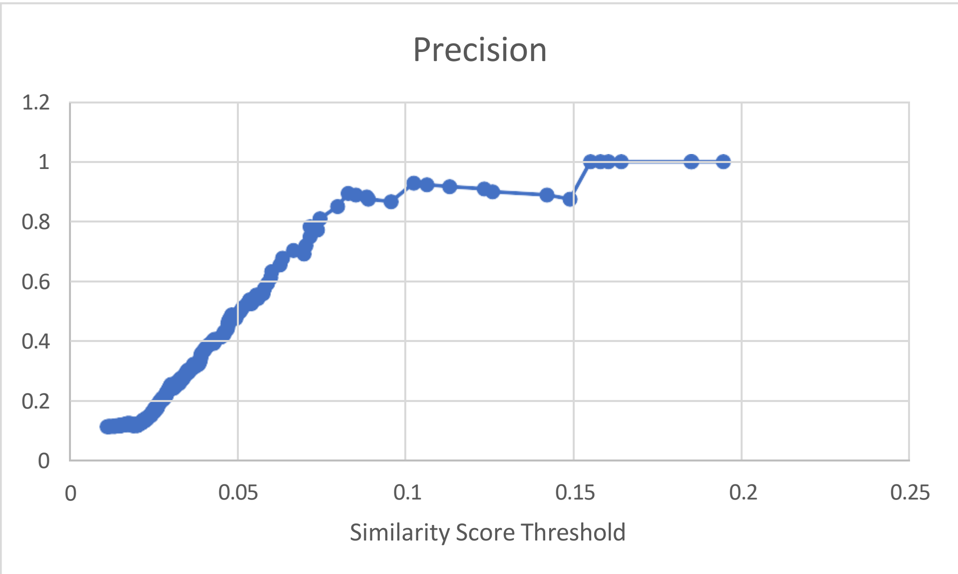

Figure 4.13 Precision of Pouring Concrete Training Dataset

doi:10.6342/NTU202401595

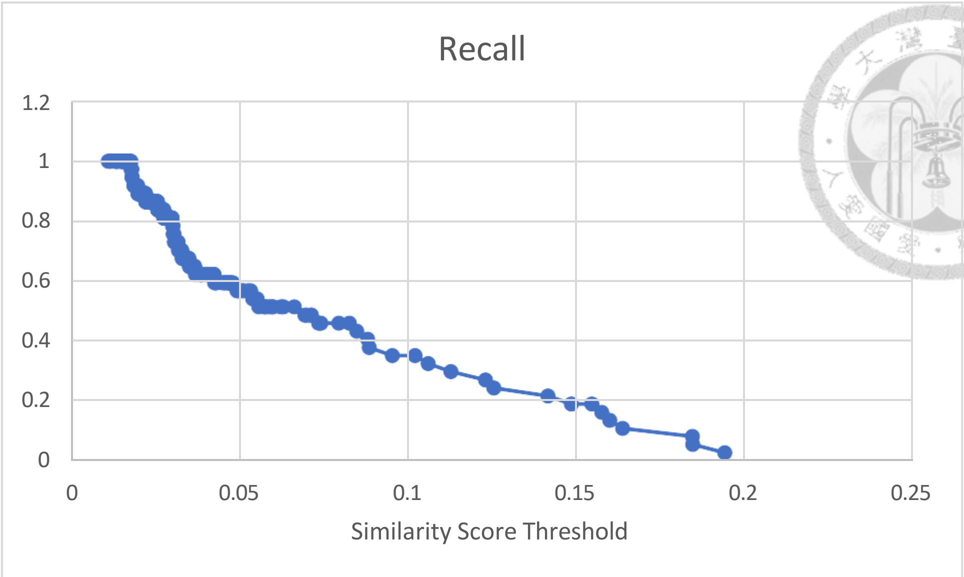

Figure 4.14 Recall of Pouring Concrete Training Dataset

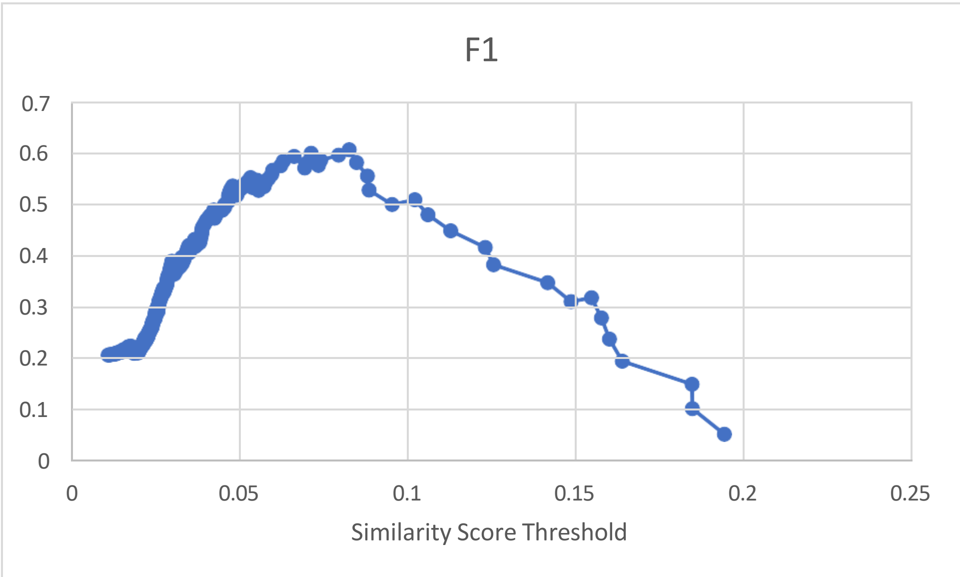

Figure 4.15 F1 of Pouring Concrete Training Dataset

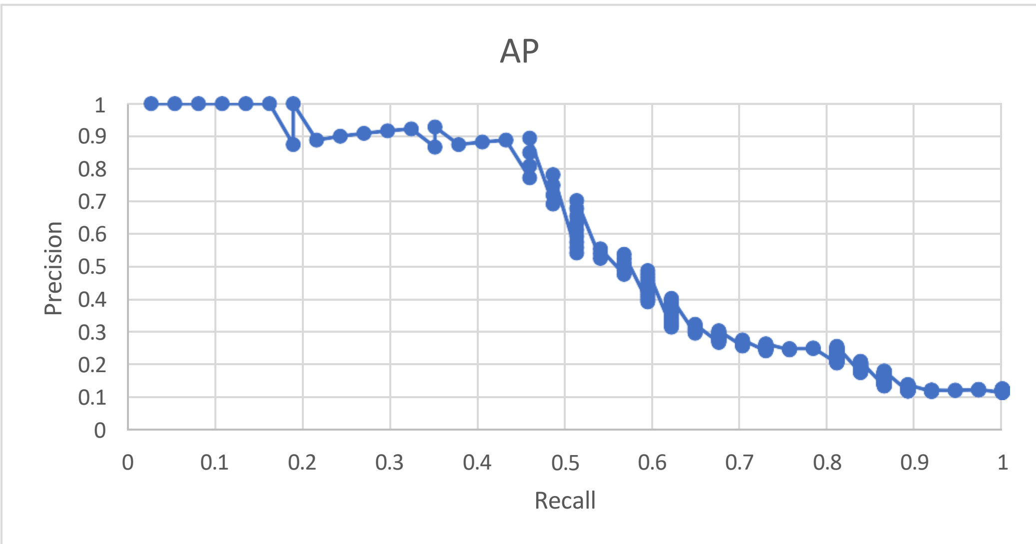

Figure 4.16 Average Precision for Pouring Concrete Training Dataset

doi:10.6342/NTU202401595

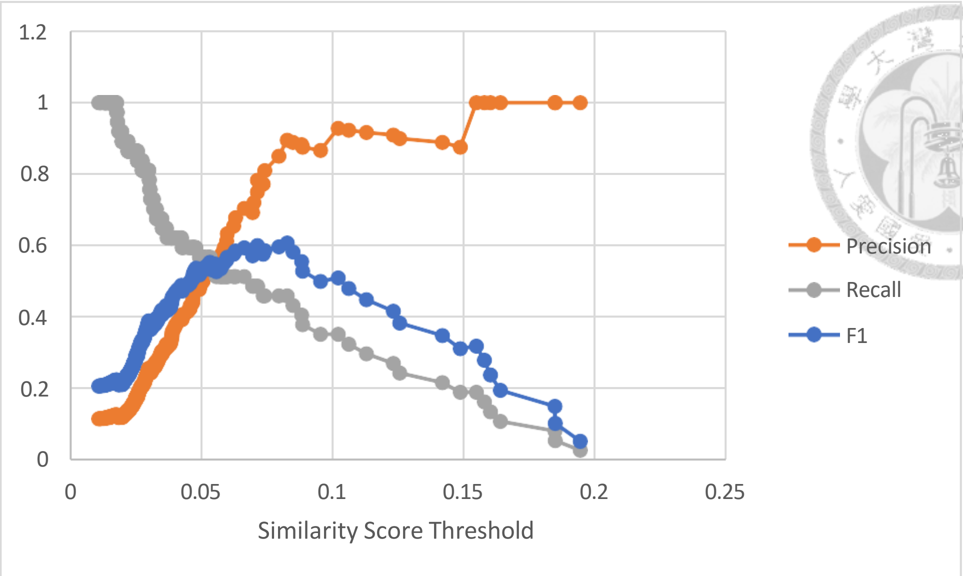

Figure 4.17 Summary of Pouring Concrete Training Dataset

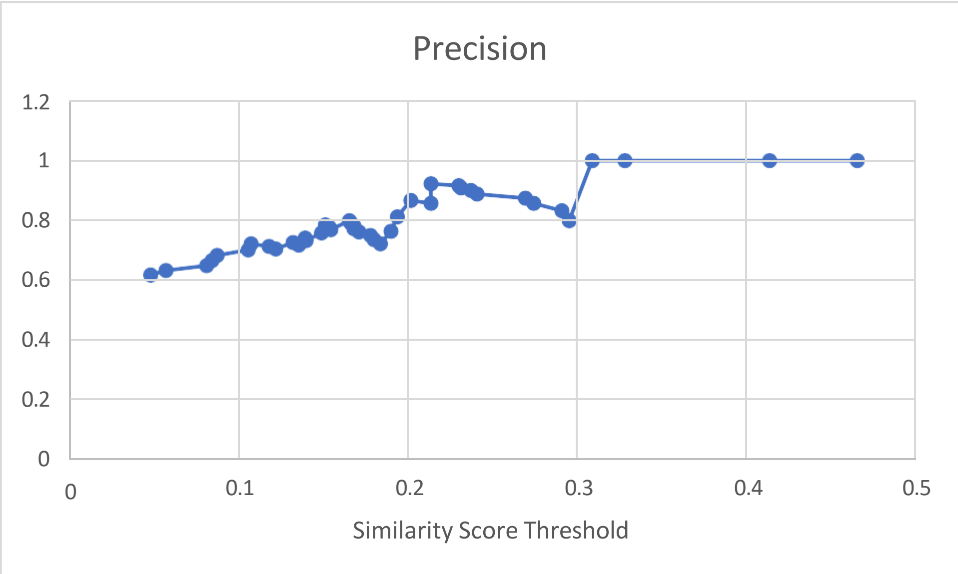

Figure 4.18 Precision of Excavation Training Dataset

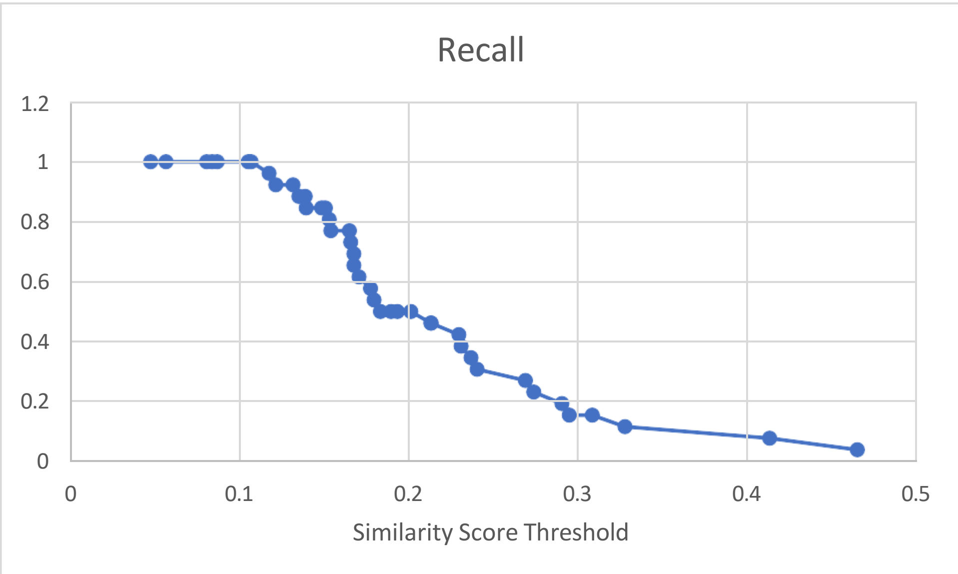

Figure 4.19 Recall of Excavation Training Dataset

doi:10.6342/NTU202401595

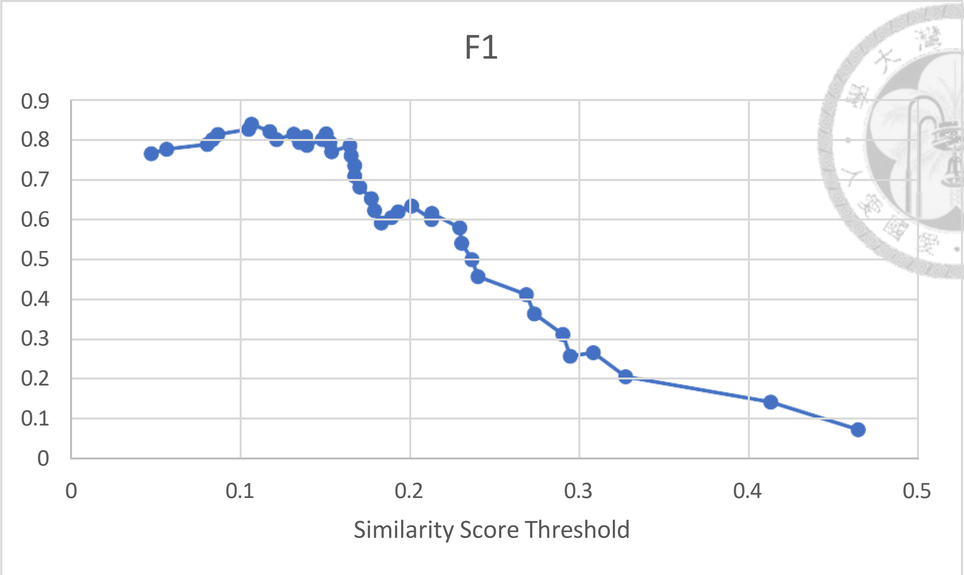

Figure 4.20 F1 of Excavation Training Dataset

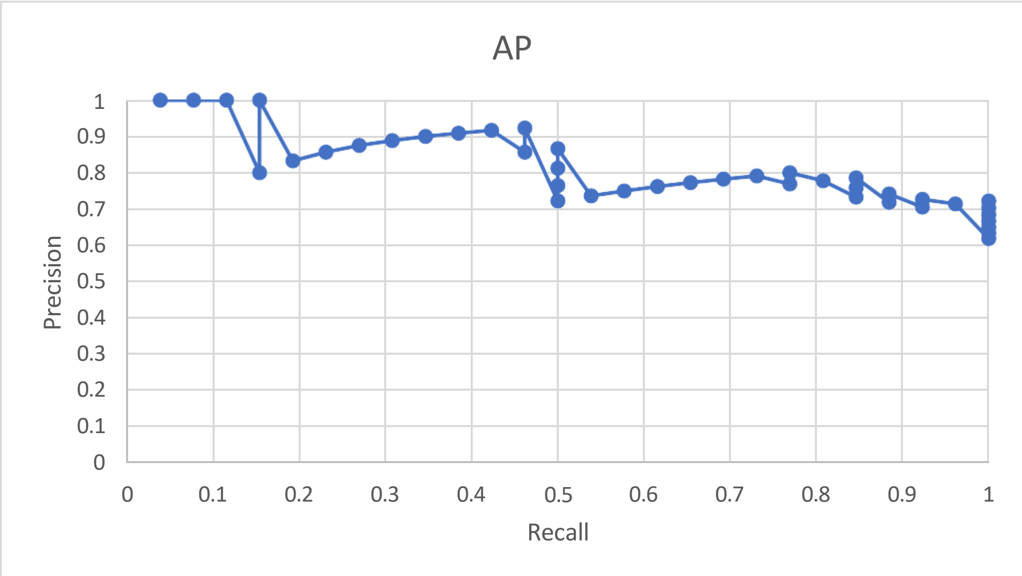

Figure 4.21 Average Precision for Excavation Training Dataset

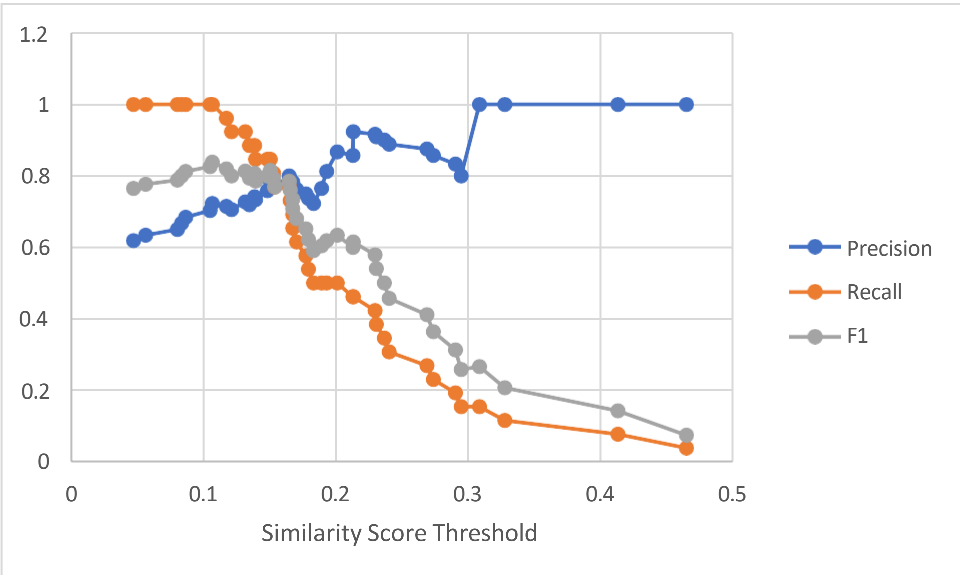

Figure 4.22 Summary of Excavation Training Dataset

doi:10.6342/NTU202401595

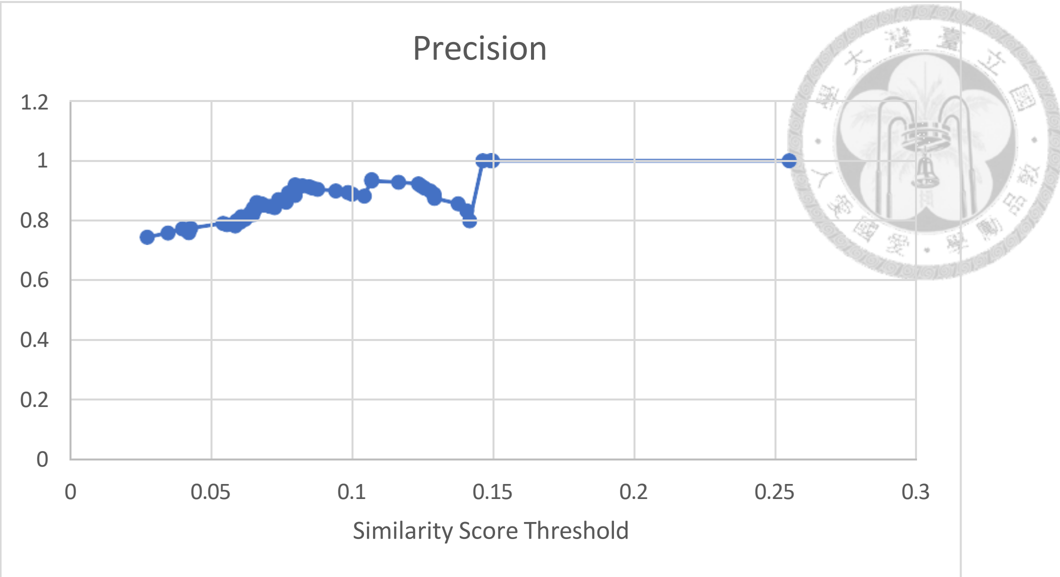

Figure 4.23 Precision of Rebar Training Dataset

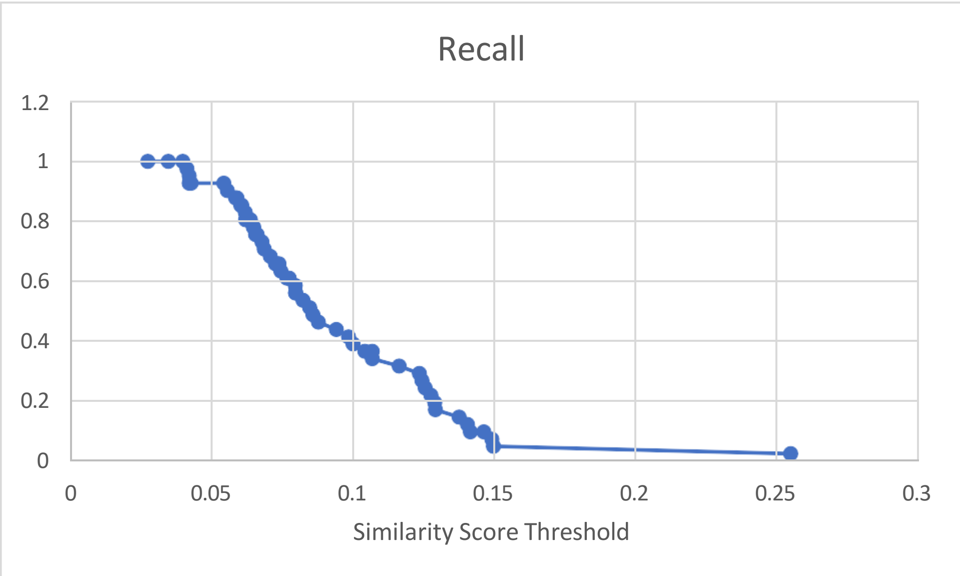

Figure 4.24 Recall of Rebar Training Dataset

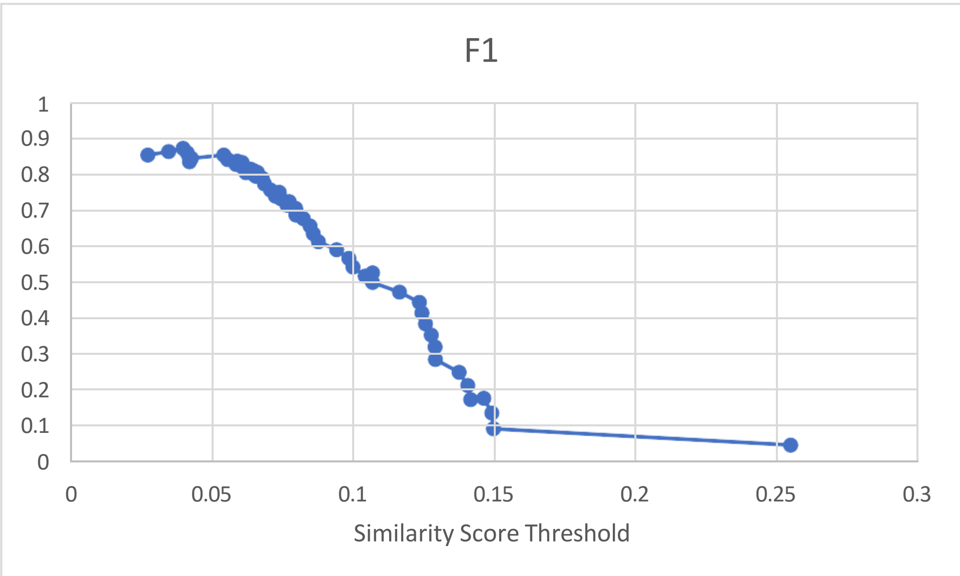

Figure 4.25 F1 of Rebar Training Dataset

doi:10.6342/NTU202401595

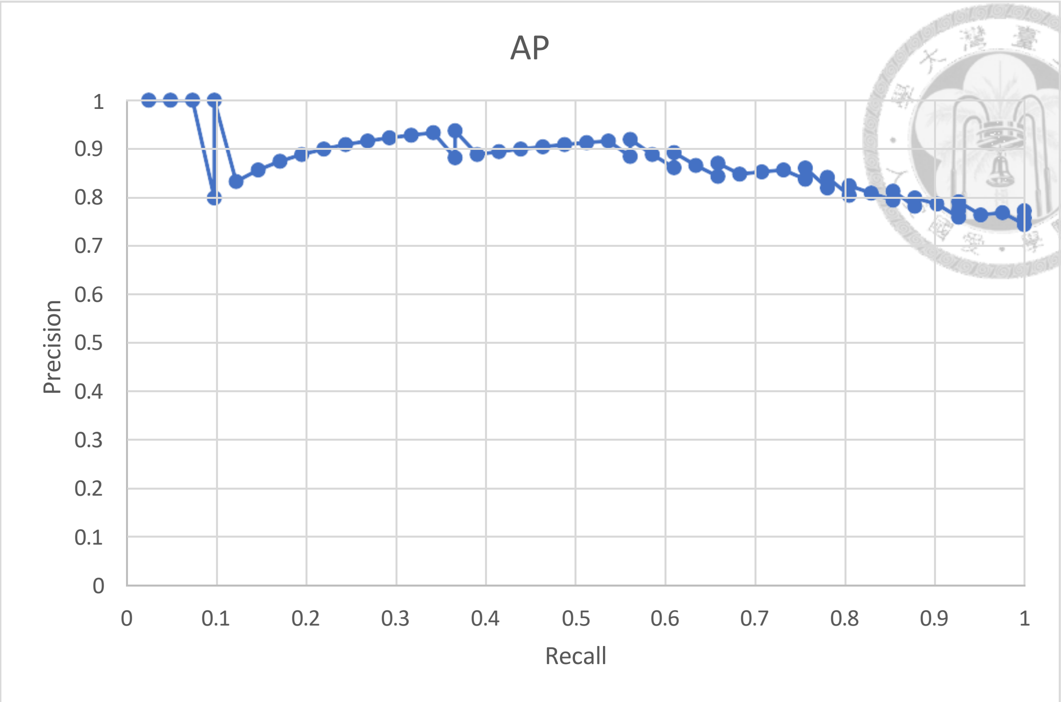

Figure 4.26 Average Precision of Rebar Training Dataset

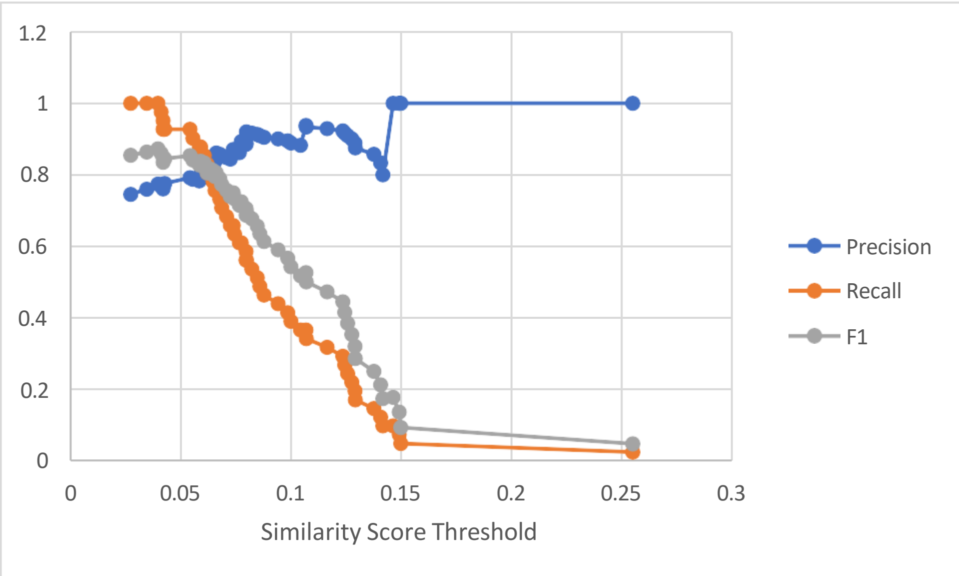

Figure 4.27 Summary of Rebar Training Dataset

Based on Figure 4.8 to Figure 4.27, it can be seen that using TF-IDF produces results that are acceptable in terms of precision, recall, and F1. The threshold functions as a similarity score filter to filter out the hazard data that are irrelevant to the activities in the schedule. It can be seen that higher similarity scores produce more relevant hazard data. But, because this research focuses more on listing all the possible hazards based on historical accident data, recall is the most important aspect to consider while precision and

doi:10.6342/NTU202401595

F1 should be kept in an acceptable value so that the result can be accurate. After finishing all the training procedures, the same steps will be applied to the 20% testing dataset. The purpose of testing the dataset is to ensure that if the data gets updated in the future, the precision, recall, and F1 score will still be acceptable. The training results can be seen in Table 4.10.

Table 4.10 Data Training and Testing Summary

Keywords	**Similarity Score Threshold**	**Item**	**Score Training**	**Score Testing**
formwork, form, formworks, forms	0.054608605	Precision	0.857142857	0.9375
		Recall	1	0.967741935
		F1	0.923076923	0.952380952
		AP	0.9092241	0.932732712
pouring concrete, pouring, concrete pump, pump, concrete pour, pour, pumping concrete, pumping (n-gram 1,2)	0.054248869	Precision	0.540540541	0.52173913
		Recall	0.540540541	0.75
		F1	0.540540541	0.615384615
		AP	0.602953162	0.685734672
excavation, digging	0.106694044	Precision	0.722222222	0.8
		Recall	1	1
		F1	0.838709677	0.888888889
		AP	0.812023144	0.875
rebar, reinforcement, rebars, reinforcements	0.039726346	Precision	0.773584906	0.590909091
		Recall	1	1
		F1	0.872340426	0.742857143
		AP	0.860787156	0.777368383

As seen in Table 4.10, in most cases, the similarity threshold chosen produces score

doi:10.6342/NTU202401595

1 on the recall score. This means that, while the precision may not be 100% accurate, a large scope of hazards has been identified, which is the main goal of this research. As for the pouring concrete, even though the results were not as good as the other activities, they are still above 50%, which means it is still more likely to be correct than wrong. There are several reasons that may contribute to why the scores were not as high as the other keywords. The first reason may be due to the number of irrelevant final narratives that contained the word “concrete”. The second reason may be due to the format of the report, some reports may be relevant, but their writing format were too long and indirect. Just as mentioned previously, this research tried to overcome the method by inputting some keywords that has been inputted as a phrase in order to balance out the word “concrete” and other words, with this method, the relevant documents’ similarity scores are raised to higher score which contributes to acceptable results as seen in Table 4.10.

4.5 Frequency and Source Count

After determining the similarity threshold for the activities, the final narratives are then selected, the ones that are below the threshold are eliminated while the ones above are used in order to print out the frequency and source count. The set-up process for frequency and source count can be seen in Figure 4.28.

doi:10.6342/NTU202401595

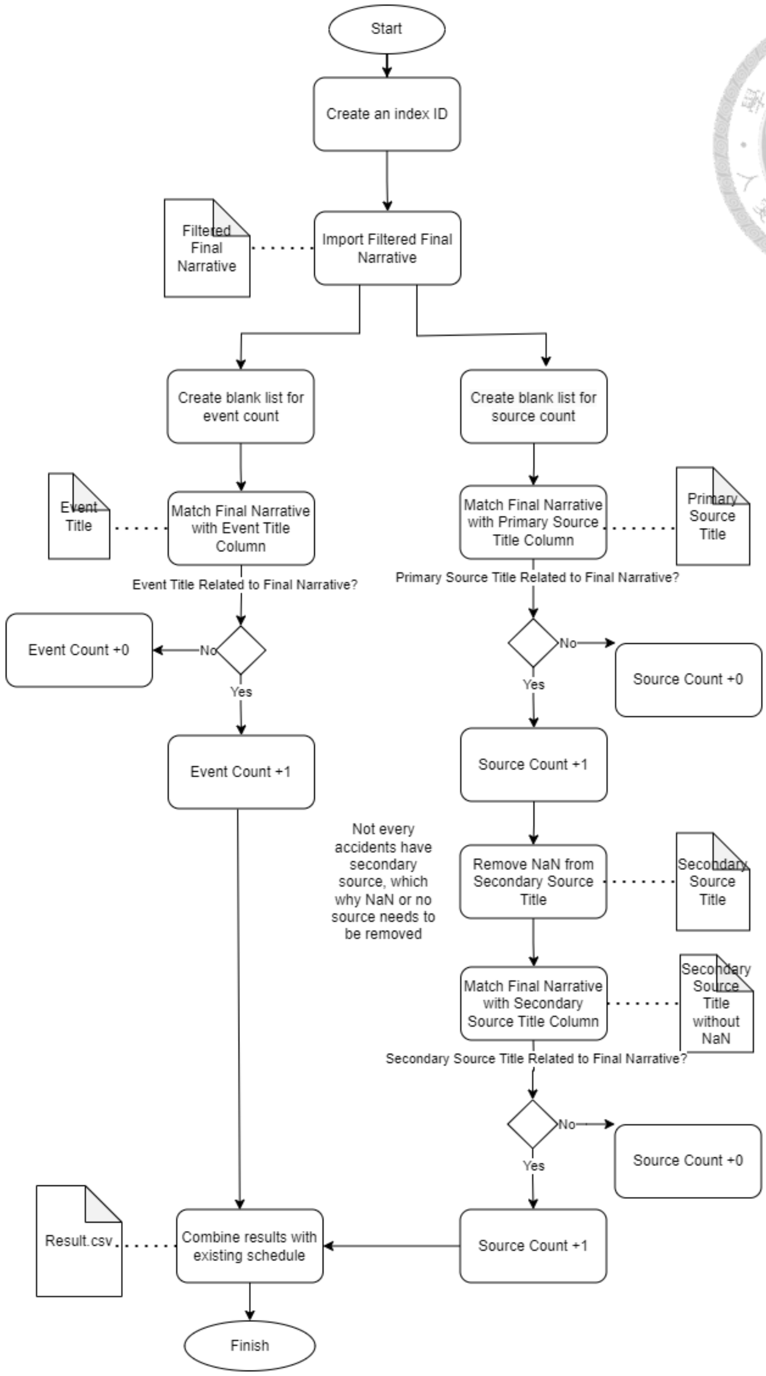

Figure 4.28 Frequency and Source Count Set Up

As seen in Figure 4.28, there are multiple steps in creating the code to print out the frequency of the hazards and the source of hazards. After checking the filtered final narrative, the program will match the result with "Event Title", "Primary Source Title", and "Secondary Source Title". The results will then be summed up and used to update the

doi:10.6342/NTU202401595

previous schedule in a csv format. Figure 4.29 shows the integration between results and schedule.

Index	Activity	Keyword	Keywords_OSHA	Hazards_sum	Source_count
3	1F-Binding formworks for w	Formwork	formwork, form, formworks, forms	{'Struck by object or equipment': 26, '	{'Structural elements, n.e.c.': 33, 'Lifeline
4	1F-Installing rebar for walls	Rebar	rebar, reinforcement, rebars, reinforce	{'Injury caused by other worker': 1, 'Ca	{'Co-worker or work associate of injured
5	1F-Casting walls	Casting	pouring concrete, pouring, concrete pu	{'Direct exposure to electricity': 1, 'Ex	{'Power lines, transformers, convertors':
6	1F-Removing formwork for v	Formwork	formwork, form, formworks, forms	{'Struck by object or equipment': 26, '	{'Structural elements, n.e.c.': 33, 'Lifeline
13	1F-Green island excavation	Excavation	excavation, digging	{'Excavation or trenching cave-in': 4, '	{'Ditches, channels, trenches, excavatio
17	2F-Binding formwoks for wa	Form	formwork, form, formworks, forms	{'Struck by object or equipment': 26, '	{'Structural elements, n.e.c.': 33, 'Lifeline
18	2F-Installing rebar for walls	Rebar	rebar, reinforcement, rebars, reinforce	{'Injury caused by other worker': 1, 'Ca	{'Co-worker or work associate of injured
19	2F-Casting walls	Casting	pouring concrete, pouring, concrete pu	{'Direct exposure to electricity': 1, 'Ex	{'Power lines, transformers, convertors':
20	2F-Removing formwork for v	Formwork	formwork, form, formworks, forms	{'Struck by object or equipment': 26, '	{'Structural elements, n.e.c.': 33, 'Lifeline
29	3F-Binding formwoks for wa	Form	formwork, form, formworks, forms	{'Struck by object or equipment': 26, '	{'Structural elements, n.e.c.': 33, 'Lifeline
30	3F-Installing rebar for walls	Rebar	rebar, reinforcement, rebars, reinforce	{'Injury caused by other worker': 1, 'Ca	{'Co-worker or work associate of injured
31	3F-Casting walls	Casting	pouring concrete, pouring, concrete pu	{'Direct exposure to electricity': 1, 'Ex	{'Power lines, transformers, convertors':
32	3F-Removing formwork for v	Formwork	formwork, form, formworks, forms	{'Struck by object or equipment': 26, '	{'Structural elements, n.e.c.': 33, 'Lifeline

Figure 4.29 Combining Results with Schedule

The final results can be seen in Figure 4.29, first the program will list all activities that contains possible hazards, then list out the keywords detected, and the keywords used for the OSHA Database, after that the possible hazards and frequencies will also be displayed along with the source of accidents and its quantity. The hazards listed and source can be seen in Figure 4.30 to Figure 4.37.

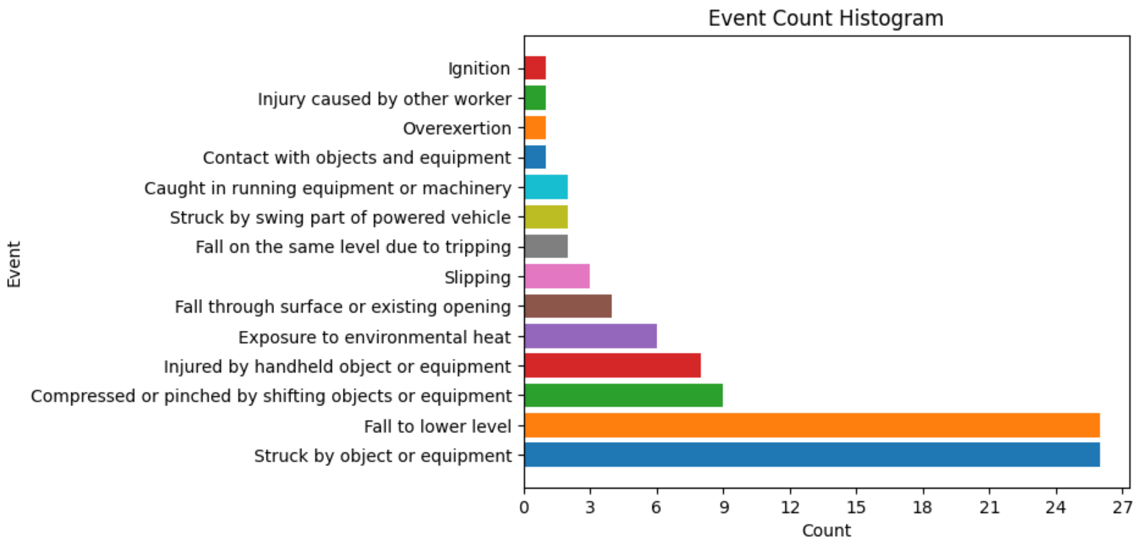

Figure 4.30 Hazard Identification for Formwork Activity

doi:10.6342/NTU202401595

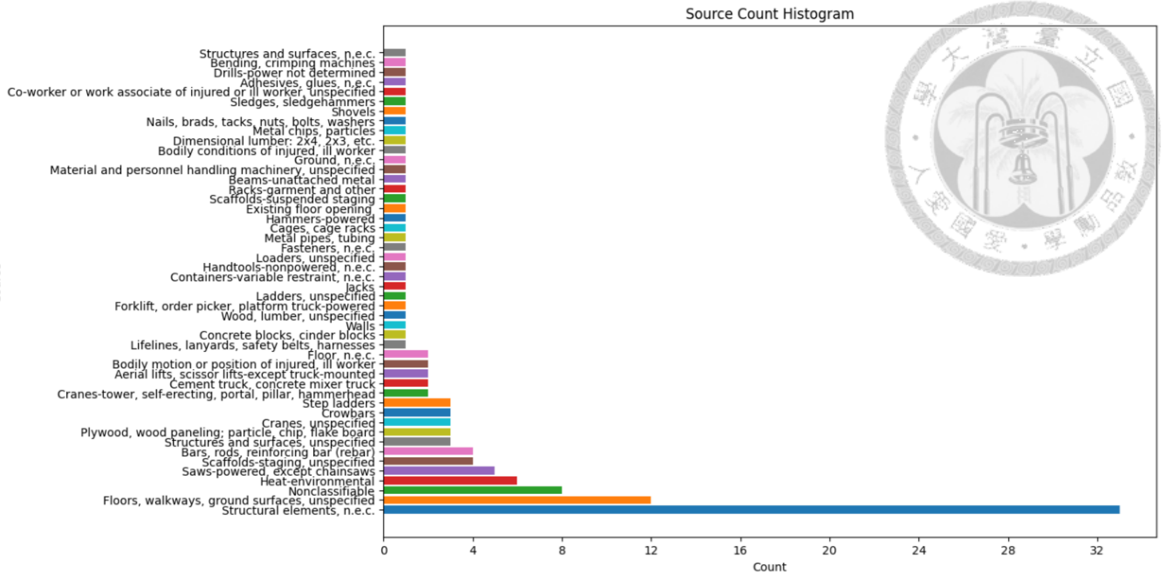

Figure 4.31 Hazard Source for Formwork Activity

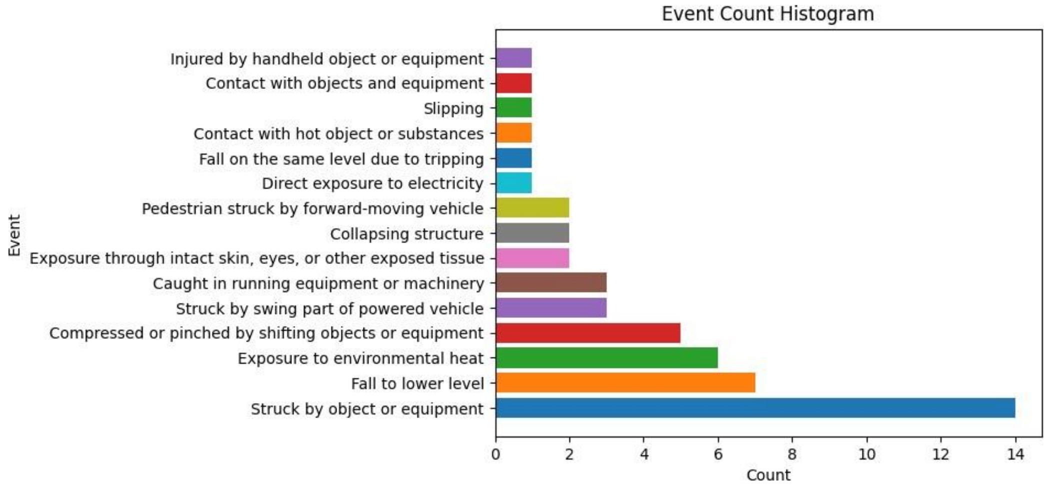

Figure 4.32 Hazard Identification for Pouring Concrete Activity

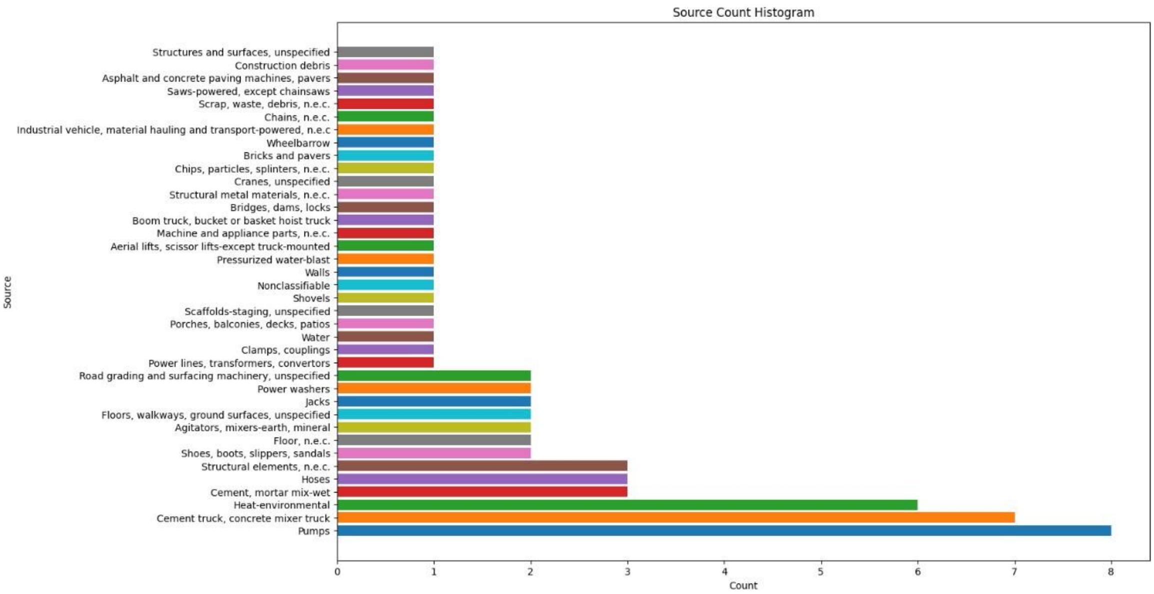

Figure 4.33 Hazard Source for Pouring Concrete Activity

doi:10.6342/NTU202401595

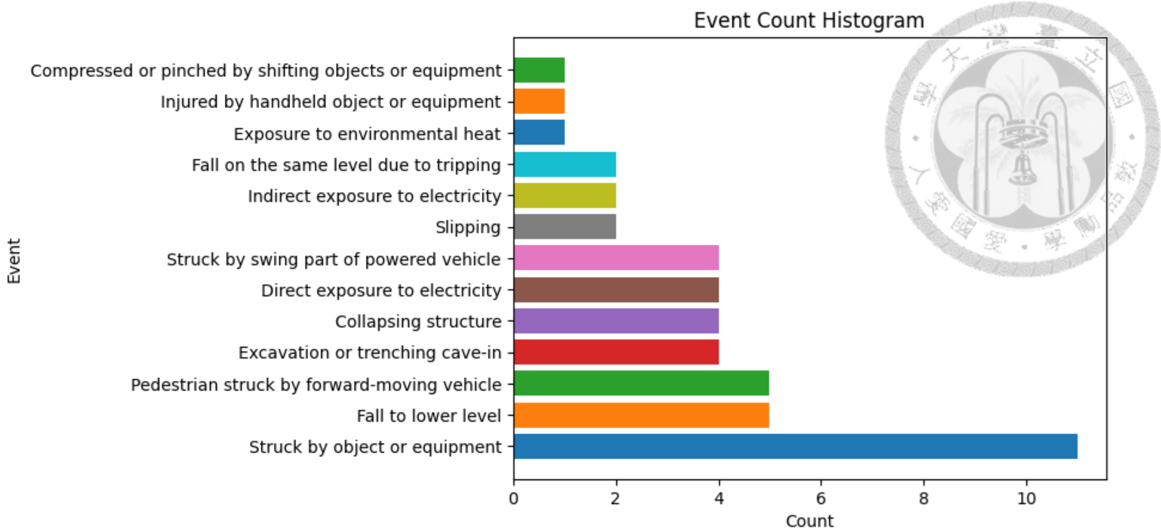

Figure 4.34 Hazard Identification for Excavation

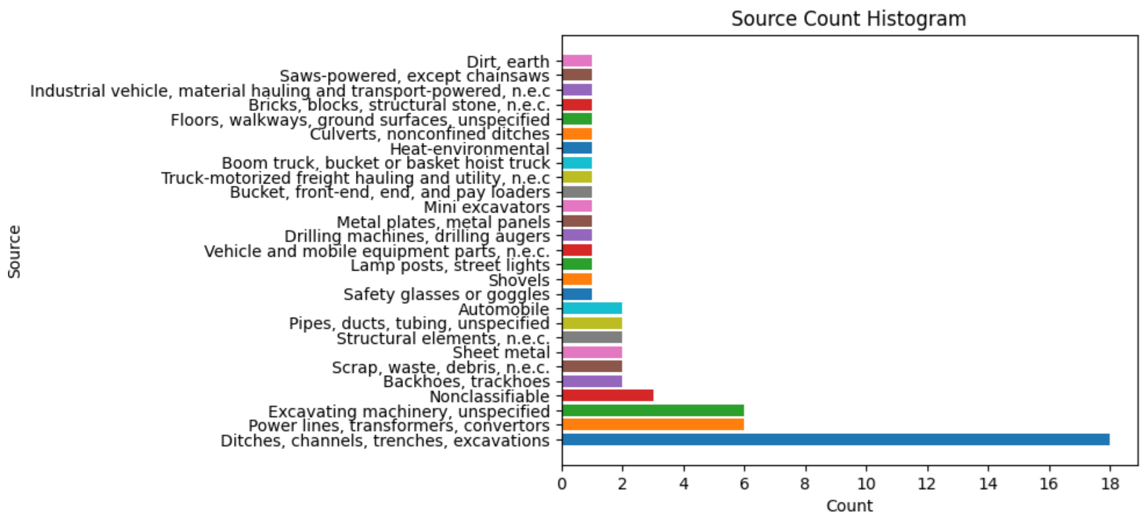

Figure 4.35 Hazard Source for Excavation Activity

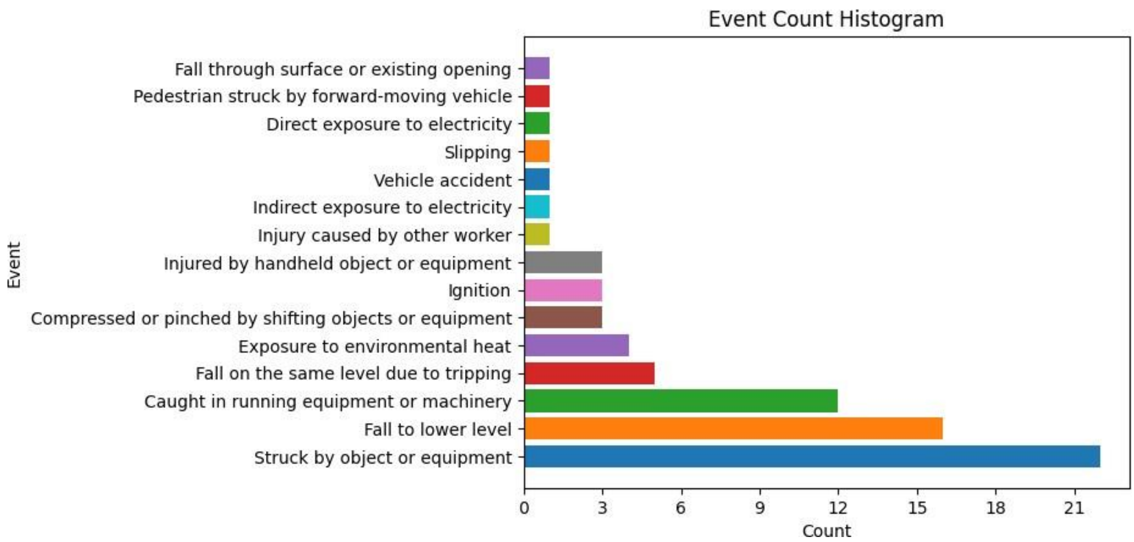

Figure 4.36 Hazard Identification for Rebar Activity

doi:10.6342/NTU202401595

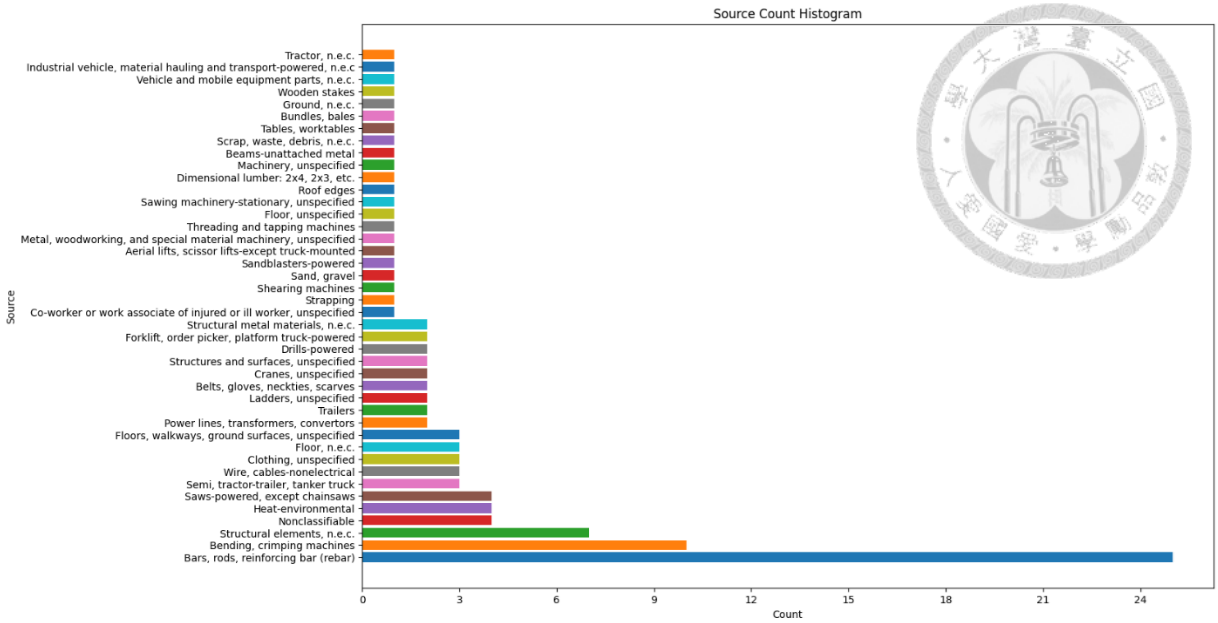

Figure 4.37 Hazard Source for Rebar Activity

As seen in Figure 4.30 to Figure 4.37, the types of hazards and the source of hazards differ from one activity to another, in some cases one type of hazard is much more dominant than the other. However, this does not necessarily mean that the other hazards are not to be taken seriously. The gap in frequency makes it easier to identify which hazards need to be taken care of first, however the severity must not be neglected either. Using the data obtained from the Automated Hazard Identification System, further judgement on the hazard analysis can be done.

doi:10.6342/NTU202401595

Chapter 5 Conclusion and Future Works

5.1 Conclusion

This research addresses the need for a faster and more precise hazard analysis system, one of them which is the hazard identification system. Using the TF-IDF to create a prototype of hazard identification system consisting of 4 representative and repetitive activities: formwork, pouring concrete, excavation, and rebar. The activities are identified from the schedule input, and then proceed to be sent to the OSHA Database for further hazards and source identification. The OSHA Database itself is a compilation of severe injury reports consisting of more than 80,000 rows of records from different industries such as farming industries, construction industries, electrical industries, transportation industries, etc. which then be sorted out based on the type of industries so that the results will be more precise and relevant to the goal of the research.

Using the activity keywords as an input for the OSHA Database TF-IDF, the system searches through the Final Narrative which consists of chronological order of how the accident happens. The results are then filtered out based on the similarity score threshold which has been trained and tested by using the OSHA Database that has been distributed randomly (80% training dataset to pick similarity score threshold and 20% testing dataset

doi:10.6342/NTU202401595

to validate the output from picked threshold) in order to prove that the result is valid and can be used for future updates. The filtering process will be based on the parameters that have been set up according to the needs of the schedule. The filtering process is then commenced using precision, recall, and F1 to obtain a wide scope of hazard list while still maintaining the quality of the precision. After that, the filtered final narrative is used to print out the types of hazards, frequencies, and sources of hazards. From the types of hazards, it can be seen that there are multiple hazards that happen more often than certain types of hazards (struck by, fall, etc.) and there are sources of accidents that caused more accidents than the others (structural elements, rebar, environmental heat, etc.).

In summary, while this method can be considered as one of the traditional forms of Natural Language Processing, its capabilities are not to be underestimated. Using similarity-based calculations, it is able to adapt to new data without relabeling unlike supervised machine learning. When compared to large language modeling, it may also provide better accuracy, since the current large language modeling needs a lot of fine tuning to avoid making hallucinations on the results. The fact that all the results, training or testing, have more than 50% precision recall F1 shows the capability of TF-IDF as a language modeling in order to provide hazard identification, frequency, and the source of hazards for hazard analysis process. The contribution of this automation system can be seen in Table 5.1.

doi:10.6342/NTU202401595

Table 5.1 Hazard Identification System Contribution

	Before	After
Time	"Because of the complexity and time-consuming nature of JHA, safety personnel must perform JHAs often weeks, sometimes even months [20]."	Hazard identification can be performed automatically, the result can even be expected in minutes
Accuracy	"Since their approach is manual and based on experience, the observed results are often error-prone due to subjective judgements of the decision maker. JHA is time-consuming, inaccurate and hard to keep up-to-date with changing construction schedules [21]."	Results are based on past accident records, adaptive according to schedule, up to date, and objective

5.2 Future Works

Since this research emphasizes the concept and building a system, it can be assumed that the result is a prototype. If this research were to be continued in the future, there are some suggestions that might be useful. First, although the writing rules for the OSHA Database follow a certain format, the prepared keywords can still be adjustable for better results. As seen in Table 4.10, although the results are above 50%, there is still a disparity between the pouring concrete activity and other activities, which means when prepared manually the quality of the keywords may be quite limited (too many words containing the word concrete and doesn't necessarily relevant to the schedule's activity). Other than that, it may be possible to integrate machine learning to improve the quality of the keywords used in the TF-IDF. However, hallucinations are to be avoided for proper results.

The other suggestions would be to gather more datasets from different countries because different countries may have different circumstances which may offer different results for the TF-IDF. Other than datasets from another country, job hazard analysis

doi:10.6342/NTU202401595

reports from companies can also improve the results considering the historical accident records will only cover accidents that are recorded on OSHA's severe injury report and not cover accidents that have not been recorded or have not happened in the past, integrating it will certainly improve the quality of the results, however it is important to equalize the format of different data sources with different writing format. Other methods (e.g. GPT, supervised learning, etc.) can also be applied together with the TF-IDF to obtain a better result. Integration between TF-IDF and other language models may be able to obtain a higher understanding about activities, steps, and possibly the relationship between each activity and the sequence of activities. To make a whole or complete hazard analysis, the hazard identification can also be integrated with the severity of the accidents provided there is a way to identify the severity level based on the injury report narrative.

doi:10.6342/NTU202401595

References

1. United States Bureau of Labor Statistics. (2023). Construction deaths due to falls, slips, and trips increased 5.9 percent in 2021. https://www.bls.gov/opub/ted/2023/construction-deaths-due-to-falls-slips-and-trips-increased-5-9-percent-in-2021.htm

2. Xu, Q., Xu, K. (2021). Analysis of the Characteristics of Fatal Accidents in the Construction Industry in China Based on Statistical Data. *International Job Environment Reservation Public Health*. 18(4): 2162. doi: 10.3390/ijerph18042162

3. Taiwan Occupational Safety and Health Administration, Ministry of Labor. (2023). Annual Report of Labor Inspection Statistics in 2022. https://www.osha.gov.tw/48783/48784/48845/48847/155008/post

4. Japan Industrial Safety and Health Association. (2021). Industrial Accidents Statistics in Japan (2021). https://www.jisha.or.jp/english/statistics/accidents_in_detail_2021.html#f07

5. United States Department of Labor. Occupational Safety & Health Administration (OSHA). OSHA Quick Card. Top Four Construction Hazards. https://www.osha.gov/Publications/3216-6N-06-english-06-27-2007.html.

6. United States Department of Labor. Occupational Safety & Health Administration (OSHA). Frequently Cited OSHA Standards Results. NAICS Code: 236 Construction of Buildings. Establishment Size: ALL sizes. https://www.osha.gov/ords/imis/citedstandard.naics?p_naics=236&p_esize=&p_state=FEFederal

7. Swuste, P., Frijters, A., Guldenmund, F. (2012). Is it possible to influence safety in the building sector?: A literature review extending from 1980 until the present.

doi:10.6342/NTU202401595

Safety Science, 50(5), 1333-1343. https://doi.org/10.1016/j.ssci.2011.12.036.

8. Driscoll, T.R., Harrison, J.E., Bradley, C., & Newson, R.S. (2008). The Role of Design Issues in Work-Related Fatal Injury in Australia. *Journal of Safety Research*, 39, 209-214. 10.1016/j.jsr.2008.02.024

9. United States Department of Labor. Occupational Safety & Health Administration (OSHA). Job Hazard Analysis. (2002). *OSHA 3071*. *https://www.osha.gov/sites/default/files/publications/osha3071.pdf*

10. Mohamed, E., Jafari, P., Preira, E., Hague, S., AbouRizk, S., & Wales, R. (2019). Web-Based Job Hazard Assessment for Improved Safety-Knowledge Management in Construction. *ISARC 2019*, 493-500. DOI: 10.22260/ISARC2019/0066

11. Singh, S.P., Mansuri, L.E., Patel, D.A., & Chauhan, S. (2023). Harnessing BIM with risk assessment for generating automated safety schedule and developing application for safety training. *Safety Science*, 164. https://doi.org/10.1016/j.ssci.2023.106179.

12. B. D. Shivahare, A. K. Singh, N. Uppal, A. Rizwan, V. S. Vaathsav and S. Suman, "Survey Paper: Study of Natural Language Processing and its Recent Applications," *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India, 2022, pp. 1-5, doi: 10.1109/CISCT55310.2022.10046440.

13. Gao, B. (2022). Research and Implementation of Intelligent Evaluation System of Teaching Quality in Universities Based on Artificial Intelligence Neural Network Model. *Mathematical Problems in Engineering*, 2022(2):1-10. DOI: 10.1155/2022/8224184

14. Dundar, H.B., Dundar, O., Ocal, H., and Kocer, S. (2022). Use of IoT and Wearable Technology Design Fundamentals in Healthcare Industry. *International Society for Research in Education Science (ISRES), Current Studies in Basic*

doi:10.6342/NTU202401595

Sciences, Engineering and Technology 2022, pp.296. ISBN: 987-605-81654-2-7.

15. Luo, X., Li, X., Song, X., & Liu, Q. (2023). Convolutional Neural Network Algorithm-Based Automatic Text Classification Framework for Construction Accident Reports. *Journal of Construction Engineering Management by ASCE,* 149 (12). DOI: 10.1061/JCEMD4.COENG-13523.

16. Ballal, S., Patel, K.A., and Patel, D.A. (2023). Enhancing Construction Site Safety: Natural Language Processing for Hazards Identification and Prevention. *Journal of Engineering, Project, and Production Management*. 2024, 14(2), 0014, DOI 10.32738

17. Singh, I., Goyal, G., and Chandel, A. (2022). AlexNet architecture based convolutional neural network for toxic comments classification. *Journal of King Saud University – Computer and Information Sciences 34 (2022).* 7457-7558. https://doi.org/10.1016/j.jksuci.2022.06.007

18. Das, B., Chakraborty, S. (2018). An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation. *Institute of Electrical and Electronics Engineers.* ***arXiv:1806.06407***

19. Han, J., Kamber, M., &Pei, J. (2012). Data Mining (Third Edition). *The Morgan Kaufmann Series in Data Management Systems*, 2, 39-82. https://doi.org/10.1016/B978-0-12-381479-1.00002-2

20. Wang, H.H. & Boukamp, F. (2011). Ontology-Based Representation and Reasoning Framework for Supporting Job Hazard Analysis. Journal of Computing in Civil Engineering ASCE, 25(6), 442-456

21. Zhang, S., Teizer, J., & Boukamp, F. (2013). Automated Ontology-based Job Hazard Analysis (JHA) in Building Information Modelling (BIM). Engineering, Environmental Science, Computer Science, Corpus ID: 161052372

doi:10.6342/NTU202401595