Transferring AI Model to CLEAR Program for Enhanced Lessons Learned and Best Practices Selection

NCDOT Project 2022-29
FHWA/NC/2022-29
May 2024

Arnav Jhala Ph.D., et al
Department of Computer Science
North Carolina State University

RESEARCH & DEVELOPMENT

Transferring AI Model to CLEAR Program for Enhanced Lessons Learned and Best Practices Selection

by

Arnav Jhala. Ph.D.

Colin Potts

Edward J. Jaselskis, Ph.D., P.E.

Siddharth Banerjee

at

North Carolina State University

Department of Computer Science
Campus Box 7902
Raleigh, NC 27695

North Carolina Department of Transportation
Research and Development Unit

Raleigh, NC 27699-1549

Final Report

Project: 2022-29

May 2024

Technical Report Documentation Page

1. Report No. **FHWA/NC/2022-29**	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle **Transferring AI Model to CLEAR Program for Enhanced Lessons Learned and Best Practices Selection**		5. Report Date **May 2024** 6. Performing Organization Code
7. Author(s) **Arnav Jhala, Ph.D., Edward J. Jaselskis, Ph.D. P.E., Siddharth Banerjee, Colin Potts**		8. Performing Organization Report No.
9. Performing Organization Name and Address **Department of Computer Science** **North Carolina State University** **Campus Box 7906** **Raleigh, NC 27699-1549**		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address **North Carolina Department of Transportation** **Research and Development Unit** **1549 Mail Service Center** **Raleigh, North Carolina** **27699-1549**		13. Type of Report and Period Covered **Final Report** **June 2022 to August 2023** 14. Sponsoring Agency Code **RP2022-29**
15. Supplementary Notes:		

16. Abstract

Transportation agency personnel gain valuable knowledge through their work, but such knowledge is lost if it is not documented properly after the worker leaves the organization. The risk of losing institutional knowledge is a current problem at state departments of transportation, including the North Carolina Department of Transportation (NCDOT), due to high personnel turnover. State transportation agencies have implemented knowledge repositories in the form of lessons learned/best practices databases to address this problem. However, motivating end-users to use such databases is challenging. This paper addresses this challenge through novel artificial intelligence technology whereby a neural network-based language model is implemented as part of the NCDOT's new knowledge management program, CLEAR (Communicate Lessons, Exchange Advice, Record). The CLEAR program encompasses a database of lessons learned/best practices and a website to access and search the database. The developed methodology involves training a language model on transportation construction texts and using that trained model in a novel algorithm enables users to search the CLEAR database easily. The developed language-processing model provides an easily accessible interface to suggest the most relevant CLEAR data based on the end-user's searched keywords. The model learns an inference model of construction domain-specific vocabulary extracted from various sources, such as contract documents, textbooks, and specifications, to make meaningful connections between lessons learned/best practices in the CLEAR database and project-specific knowledge. The developed model has been validated by project managers for projects at various lifecycle stages. The automation of information retrieval is intended to encourage NCDOT personnel to use and embrace the CLEAR program as part of their routine work to improve project workflow. In the long run, the NCDOT will benefit from consistent usage of the CLEAR program and the high-quality content that is input to the CLEAR database, thereby leading to enhanced institutional knowledge and organizational innovation.

17. Key Words **Artificial Intelligence, Fine-tuned Language Models, web-based lessons learned database, data visualization, information dissemination**	18. Distribution Statement		
19. Security Classif. (of this report) **Unclassified**	20. Security Classif. (of this page) **Unclassified**	21. No. of Pages **65**	22. Price

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

DISCLAIMER

The contents of this report reflect the views of the authors and not necessarily the views of North Carolina State University. The authors are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the North Carolina Department of Transportation or the Federal Highway Administration at the time of publication. This report does not constitute a standard, specification, or regulation.

ACKNOWLEDGEMENTS

The research team acknowledges the North Carolina Department of Transportation (NCDOT) for supporting and funding this project. We extend our thanks to the project Steering and Implementation Committee members:

Janaki Patel	**Chair**
Alyson Tamer	**Vice-Chair**
Stephen Morgan	**Member**
Jordan Woodard	**Member**
Thad Duncan	**Member**
Alexander Foster	**Member**
Brian Radacovic	**Member**
Curtis T. Bradley, Ph.D.	**P Member**

The authors also thank NCDOT personnel who participated in this research project for their time and hospitality. Without the help of all these individuals, the project could not have been completed in such a successful manner. The active participation and resulting contributions of NCDOT personnel and the Steering and Implementation Committee were especially noteworthy and helpful.

EXECUTIVE SUMMARY

The North Carolina Department of Transportation (NCDOT) created a new knowledge repository called Communicate Lessons, Exchange Advice, Record (CLEAR) as an official platform for end-users to store and retrieve knowledge. Through the CLEAR program, end-users can enter lessons learned and best practices gained in their workplace in addition to soliciting solutions to any ongoing issue. This project proposed to transfer an artificial intelligence (AI) model using natural language processing that improves the search capabilities of the CLEAR Program to more efficiently identify relevant lessons learned and best practices. The CLEAR Program includes a collection of documented lessons learned and best practices primarily entered in the form of text with some image files. A construction language inference model has been developed that can make meaningful connections between lessons learned, best practices, and construction domain vocabulary (e.g., a fiber optic cable would be recognized as a utility in the AI model). A proof-of-concept AI model will be validated by project managers on a set of pre-selected projects whose information will be obtained from the NCDOT Value Management Office. This validation will certify the usefulness of the generated keywords and thereby the AI model in an iterative manner until the model has been appropriately fine-tuned. We integrated the language model into the value management office process to record, update, and improve their QA/QC guidelines. This project serves as a pilot project to demonstrate the generality and usefulness of the model across departments within the NCDOT.

We have documented the process to make it easier for other departments to consider the integration of this language model. In addition, this effort focused on the transference and implementation of the AI model on NCDOT servers. In the long run, this automation in information retrieval will encourage NCDOT personnel to use the CLEAR program as a part of their routine work to improve project workflow processes. By storing and retrieving knowledge for future projects, this repository will help the NCDOT to achieve better project control and to be better prepared to consider suggestions for innovative ideas, thereby adding value to the state of North Carolina.

To ensure CLEAR's proper functioning and maximum reach for NCDOT personnel, this research utilized cutting-edge concepts of artificial intelligence (AI) and data visualization to encourage the process of knowledge sharing. A data dashboard tailored for the gatekeeper provides effective means to monitor progress that relates to predetermined metrics. The dashboard serves as a success metric for the CLEAR program by monitoring entries based on factors such as the status of implementation of various lessons and best practices, Innovation Culture Index survey data to assess end-users' ability to innovate, and website analytics data developed in a previously funded project. The AI-enabled set of language embeddings fine-tuned to construction vocabulary helps provide useful insights about the text that is entered into the knowledge repository by effectively disseminating information, thus allowing the utilization of wisdom within the knowledge repository to be a proactive process.

The final research products are (1) a comprehensive lessons learned/best practice resource repository that can be used to improve performance for future NCDOT projects, (2) a data dashboard to enable the gatekeeper to monitor the progress of the end-users and intervene when necessary, and (3) an AI-based model to disseminate information to end-users automatically. The NCSU research team has provided these products to the NCDOT Value Management Office in conjunction with a presentation that includes a demonstration of the dashboard and AI model to ensure that these products are in line with increased end-user participation in the CLEAR program. The dashboard and AI model are envisioned to provide useful insights and automatically disseminate relevant information that is best suited to stakeholders' needs. The NCDOT will greatly benefit from the language model program and database as well as from applications of the data analysis-enabled products, thereby improving project management and operational performance for the long term.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Construction projects are dynamic and seldom repetitive in nature, unlike other sectors such as manufacturing that typically have a routine set of tasks to create a product. The architecture, engineering, and construction industry employs predominantly project-based teams whose members have various levels of work experience and knowledge. Personnel accrue incremental knowledge from many projects over their careers. From a construction project's inception to its handover, construction project teams tend to work interdependently on various project lifecycle phases comprising design, construction, and maintenance. Such project variation and personnel interdependence lead to a learning curve that requires additional training, which in turn consumes additional project resources, i.e., time and money (Johari & Jha, 2021; Everett & Farghal, 1994).

As an intangible asset, experiential knowledge is difficult to associate with a direct monetary value (Dekker & de Hoog, 2000), yet it is a valuable asset to an organization as long as the employee remains employed. However, upon their departure from the organization, whether due to personal reasons or retirement, many years' worth of knowledge is lost if the information is not properly recorded or stored. That is, team turnover or employee retirement can lead to a huge loss of institutional knowledge. Currently, most state departments of transportation (DOTs) are facing excessive personnel turnover rates, with most state DOTs reporting 10-12% annual turnover rates (McRae, Vallet and Jewiss 2018). Often, the departure of a particular team member creates a void in the key skill sets that the remaining team members find difficult to fill within a short time span. The struggle to retain existing personnel and train fresh recruits can lead to the need to allocate additional resources toward developing strategies to keep employees motivated (AASHTO Journal, 2021). To compound the turnover problem, more than half of the current DOT workforce will be eligible for retirement in the next five years (National Skills Coalition, 2021). Within the span of the current decade, all baby boomers (people born from 1946 to 1964) will be at least 65 years of age, or in other words, past the conventional retirement age of 62 years. Furthermore, the recent COVID-19 pandemic has exacerbated the already worsening construction workforce shortage, requiring organizations to take additional measures to ensure project continuation despite team turnover (Alsharef A. , Banerjee, Uddin, Albert, & Jaselskis, 2021; Assaad & El-adaway, 2021). The aforementioned factors can lead to negative

impacts on the quality of work and key project metrics such as schedules and budgets. To combat these negative impacts from an organizational perspective, harnessing previously acquired knowledge can help reduce repeated mistakes and improve organizational efficiency for project delivery (Amir & Parvar, 2014).

To ensure that the knowledge gained by construction personnel will remain within the organization, various knowledge management techniques are now available to organize and store knowledge in the form of checklists or databases. This collective knowledge is termed 'organizational capital', which becomes proprietary to the organization and thus can be used even after individuals have left the organization (Youndt & Snell, 2004). That is, the organization owns and controls the knowledge and is not beholden to or dependent on any specific individual within the organization, thereby providing a stable and consistent knowledge base over time. A learning organization is one that facilitates quick and effective knowledge transfer among project team members via knowledge management repositories and in which such databases are widely adopted in the realm of organizational knowledge management and innovation (Goh, 2002). In addition to providing an official platform for collaborative team learning, knowledge repositories can help organizations promote internal organizational innovation. This effort also leads to reducing repeated mistakes and achieving enhanced project outcomes, thereby improving the organization's competitive edge in the market (Ferrada, Nunez, Neyem, Serpell, & Sepulveda, 2015).

Despite having operative lessons learned/best practices databases in place, organizations still struggle to reap the full benefit of such knowledge repositories. The biggest success factors for ensuring that knowledge repositories justify their purpose are (1) end-users' willingness to embrace such databases as part of their routine work and (2) end-users' ability to store and retrieve knowledge at will. In addition to coping with their regular job responsibilities, end-users do not necessarily have the time or inclination to carve out extra time from their work schedule to peruse knowledge in a database (Fullerton C. E., Tamer, Banerjee, Alsharef, & Jaselskis, 2021). Failure to motivate end-users to use a knowledge repository will ultimately render the repository defunct, and thus, the efforts to create the repository are rendered futile as well. Moreover, at an organizational level, obsolete knowledge repositories can lead to repeated mistakes and diminished internal innovation, causing financial-related problems for the

organization. From an information systems (IS) perspective, collaborative end-users and supportive upper management are critical for ensuring long-term success of organizational IS such as knowledge repositories (Petter, DeLone and McLean 2013). Besides, knowledge repositories are also effective in promoting internal organizational innovation, which aids in preserving market competitiveness by institutionalizing knowledge (Zahra and George 2002). Thus, proactively providing a mechanism for end-users to use knowledge from the repository and stay engaged with the process is needed to ensure that personnel contribute to and utilize high-quality database content.

This project developed a novel set of algorithms in the context of enhancing end-user usage of an already established construction knowledge repository CLEAR developed by the NC DOT Value Management Office. The resulting AI model is referred to as the Construction Domain-Specific Artificial Intelligence Language (CD-SAIL) model. The CD-SAIL model, which involves natural language processing, is designed to identify lessons learned/best practices automatically and intelligently based on keyword(s) entered by end-users. Specifically, the CD-SAIL model makes meaningful connections between the entered keyword or phrases and the existing lessons learned/best practices stored within the North Carolina Department of Transportation's (NCDOT's) new knowledge management program called CLEAR (Communicate Lessons, Exchange Advice, Record). The CLEAR program includes a knowledge database and a website to access and search it. The objective of the developed methodology is to provide automated information retrieval that encourages NCDOT personnel to use the CLEAR program as part of their routine work to improve project workflow, which in turn should benefit the NCDOT with enhanced institutional knowledge and organizational innovation.

2. LITERATURE REVIEW

Tacit and explicit knowledge

Managing construction projects involves coordinating different stakeholders across the project lifecycle phases that include planning, design, construction, operations, and maintenance. During project execution, project staff members gain valuable knowledge, experience, and lessons learned and then apply their accumulated knowledge to future projects (Nonaka, 1994). Such acquired knowledge typically is manifested as either explicit knowledge, which can be communicated clearly through formal systematic language, or tacit knowledge, which is deeply rooted in action

and commitment. Formal systematic language is difficult to communicate (Smith M. K., 2003; Nonaka, 1994). More than 80% of all the knowledge gained by construction personnel can be classified as tacit knowledge and the remaining as explicit knowledge (Sheehan, Poole, Lyttle, & Egbu, 2005). Although explicit knowledge can be easily documented and reused by personnel, tacit knowledge can be difficult to store and retrieve, especially as such knowledge is deeply embedded and sometimes difficult to express in words. Addis (2016) notes that it can be difficult for construction personnel to convey tacit knowledge gained on project sites during their routine work. Despite being difficult to accomplish, converting tacit knowledge so that it can be expressed more easily in explicit terms is a worthwhile effort. Codification or other means can be used to ensure the smooth transfer of knowledge from one person to another to institutionalize knowledge within the organization. With the advent of the latest innovations in information technology, organizations are now able to use digital formats to store and retrieve knowledge, both tacit and explicit, in the form of lessons learned databases that capture knowledge using a set of rules and are vetted by experts to ensure high-quality input into such repositories (Anumba, Egbu, & Carrillo, 2005; Egbu, 2004).

Lessons learned/Best practices databases.

Lessons learned databases have proven to be effective organizational tools to store and retrieve past knowledge (Rowe and Sikes 2006). The Project Management Institute (2017) defines lessons learned as the learning gained from the process of performing the project. Professional organizations also are creating standard frameworks for preserving project knowledge. For example, the Construction Industry Institute lists lessons learned as one of its 17 best practices to facilitate the continuous improvement of organizational processes and procedures by institutionalizing knowledge. The lessons learned process is comprised of three steps:

- Collection: End-users identify specific problem areas to be analyzed as lessons learned.
- Documentation: The identified lessons are documented using a formal mechanism, most commonly in the form of a lessons learned database.
- Communication: Documented lessons are communicated to the people who could benefit the most by gleaning these knowledge ‘nuggets.’

A successful lessons learned program is one where end-users are able to make use of all three steps in their future projects. Another important factor that impacts the success of lessons learned

programs is the willingness of end-users to embrace such programs by entering high-quality content into the database or searching for stored knowledge to be applied to future projects. This willingness of end-users to take advantage of lessons learned is a critical consideration that many organizations fail to address, ultimately leading to a failed knowledge management program. Institutionalizing knowledge can help organizations promote internal innovation, reduce repeated mistakes, and maintain market competitiveness by having efficient and improved workflow.

Knowledge representation in the construction industry

Ontologies provide a shareable mechanism for classifying domain knowledge and facilitating the semantic exchange of knowledge. Specifically, ontological web languages (OWLs) have gained popularity by using semantic web-based technologies to facilitate dynamic information-sharing. Being knowledge-intensive in nature, the construction industry requires project teams for various project lifecycle phases to work collaboratively and share knowledge that is gleaned during all project phases for enhanced project outcomes (Issa & Haddad, 2008). This effort generally is accomplished using OWLs that are created using construction-specific domain terminology rather than general dictionary-based semantic terminology.

Expert domain knowledge can be represented using either rule-based reasoning or case-based reasoning. In rule-based reasoning, the computer system emulates the decision-making ability of human experts based on the knowledge within the domain ontologies, generally by using if-then rules in a deductive way. That is, satisfying the conditions of a rule will lead to some conclusion or an action being performed. In a case-based reasoning mechanism, the computer system is fed a set of historical or theoretical prototype problems and their solutions in an inductive manner. In a case-based system, new problems are solved by analogy, which is matching and adapting cases that previously have been solved successfully (Berka, 2011).

Representing domain knowledge: Current AI approaches

Construction involves many unstructured or semi-structured text documents that are written in natural language. Many researchers have attempted to leverage the tacit knowledge contained in such documents using various levels of automation via natural language processing. For example, Rezgui (2006) used an ontology-based approach to summarize documents for information retrieval. Several other rule- and ontology-based approaches include the manual creation of knowledge map models (Tserng, Yen-Liang, & Lee, 2010) and the ruled-based extraction of risks

from construction contract documents (Lee, Yi, & Son, 2019). Identifying important information from documents also includes automated compliance checking, which incorporates semantic rule-based extraction using domain ontology (Zhang & El-Gohary, 2016), and deep-learning approaches such as bidirectional long short-term memory neural networks (Zhang & El-Gohary, 2021). Beyond extracting risks from contract documents, Hassan and Le (2020) used an assortment of machine learning techniques (including naive Bayes, support vector machines, logistic regression, and feedforward neural networks) to find text that indicates requirements in contract documents. Similar to the work undertaken in our current research, Kim and Chi (2019) investigated searching and extracting important information from accident case reports. They used Okapi BM25 (a search-ranking algorithm), rule-based extraction, and conditional random fields to yield impressive search results. In addition to information extraction and searches, Kim et al. (2022) used bidirectional encoder representations from transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2019) to build a question-answering system to find infrastructure damage information. This area of research overall has seen significant gains using machine learning techniques for natural language processing.

Many frameworks are available for language models that have applications for information retrieval (Zhai, 2008). A recent development is neural network language models that embed the high-dimensional space of words into a relatively low-dimensional continuous vector space. These models include Doc2Vec (Le & Mikolov, 2014), GloVe (Pennington, Socher, & Manning, 2014), and FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017). Other competing models include BERT (Devlin, Chang, Lee, & Toutanova, 2019), which is used by Google in its search engine, and combined techniques such as spaCy (Honnibal, Montani, Van Landeghem, & Boyd, 2020) that can use BERT and other models internally. All of these frameworks define unique models that support many different tasks involving natural language. Viewed as language models, they encode an understanding of the target language in the training dataset. Thus, these frameworks can either be general models, such as Wikipedia or other large non-specialized sources that are trained in a common language, or a domain-specific corpus that results in the trained model not having a general understanding of the language but an understanding that is specific to the target domain. Prior work using Doc2Vec for transportation construction text highlights the effectiveness of learning vocabulary obtained from transportation construction texts (Banerjee S. , Potts, Jhala, &

Jaselskis, In press). Some techniques, such as BERT, often are first trained on a large general corpus and then specialized for a particular domain.

3. NCDOT's CLEAR Program

The NCDOT has witnessed high rates of personnel turnover over the past two decades, which has been one of the most significant contributors to project delays (NCDOT, 2004; Fullerton C. E., Tamer, Banerjee, Alsharef, & Jaselskis, 2021) and loss of institutional knowledge. Moreover, until recently, NCDOT personnel had no official platform to communicate the knowledge gained in their routine work, which led to a silo mentality with no means to reach out to a broad audience. The most common way to learn about a new technique or materials that either worked well or did not work well on a project was through word-of-mouth during conferences or telephone calls. Additionally, project teams from different project lifecycle phases often fail to convey feedback to personnel working on other phases, thereby repeating the same mistakes in the absence of correction mechanisms. Poor communication of project information among project teams and the increase in repeated mistakes negatively affect project outcomes in the form of increased change orders and supplemental agreements for the NCDOT (NCDOT, 2004). These additional project costs and time could have been mitigated by implementing an official platform to share project knowledge.

To address the need for such a knowledge-sharing platform, the Value Management Office (VMO) at the NCDOT conducted a preliminary study in 2014 titled 'Post-Construction Assessment Program.' This study aimed to understand common problems faced by project teams once the project reached the substantial completion stage. A simple interview guide with open-ended questions was used to solicit responses from various project team personnel in hydraulics, photogrammetry, geotechnical, construction, and maintenance regarding their most recent project experiences. The survey responses indicated the need to provide an official platform for personnel to communicate their experiences at all project lifecycle phases for improved project delivery outcomes. The identified need for a communications platform led to the conceptualization of a new knowledge repository program (with an inherent database and website) called CLEAR (Communicate Lessons, Exchange Advice, Record) in fall 2017 in consultation with researchers from North Carolina State University. The CLEAR program provides opportunities for end-users to access and share lessons learned/best practices with the goal of improving their ability to solve project-related problems and improving project workflow. The sole beneficiaries of the CLEAR

program are NCDOT personnel and project teams that are involved with the construction and maintenance of infrastructure projects sponsored by the NCDOT in the state of North Carolina.

Principal CLEAR program stakeholders

The principal stakeholders involved with the CLEAR program, and their responsibilities as they pertain to CLEAR, are as follows:

End-users: End-users are NCDOT personnel who are responsible for entering useful lessons learned and best practices based on knowledge gained at project sites. They are also responsible for searching for relevant knowledge to understand previous circumstances to avoid repeating mistakes.

Gatekeeper: The Value Management Office (VMO) at the NCDOT serves as the gatekeeper for the CLEAR database and is responsible for checking the completeness of lessons learned/best practices submissions, forwarding the submissions to taskforce members, and subsequently approving the submissions after receiving the go-ahead from taskforce members.

Taskforce: The taskforce is composed of experts from various disciplines responsible for ensuring the quality of the content uploaded to the database. Based on its review of each submission, the taskforce informs the gatekeeper of its decision to accept or reject the submission. Note that, whereas the taskforce consists of experts who cover all disciplines of work, Expert Review Panel members are selected by the gatekeeper from this pool of experts as those who can offer the most relevance and expertise to evaluate the submissions.

Innovation Coordinators: These coordinators are highly motivated personnel who encourage their units or offices to participate in the CLEAR program, thereby supporting innovation.

Technical Advisory Group: This group is composed of taskforce/Expert Review Panel members who focus on specific topics or areas and collectively review submissions and establish goals for solutions.

Technical Coordination Committee: This committee is composed of upper management, multi-disciplinary and multi-modal representatives, and external partners who provide guidance and review from a high-level industry perspective.

Figure 1 shows the progression of a lessons learned entry within the CLEAR portal and the roles of various stakeholders in this process. The flowchart was color-coded to describe the major

stakeholders' roles throughout the process. Usage model can be an effective tool for generating multimedia summaries (Potts & Jhala, 2021) and for interactive summarization of data filtering and triage (Robertson, Harrison, & Jhala, 2020).

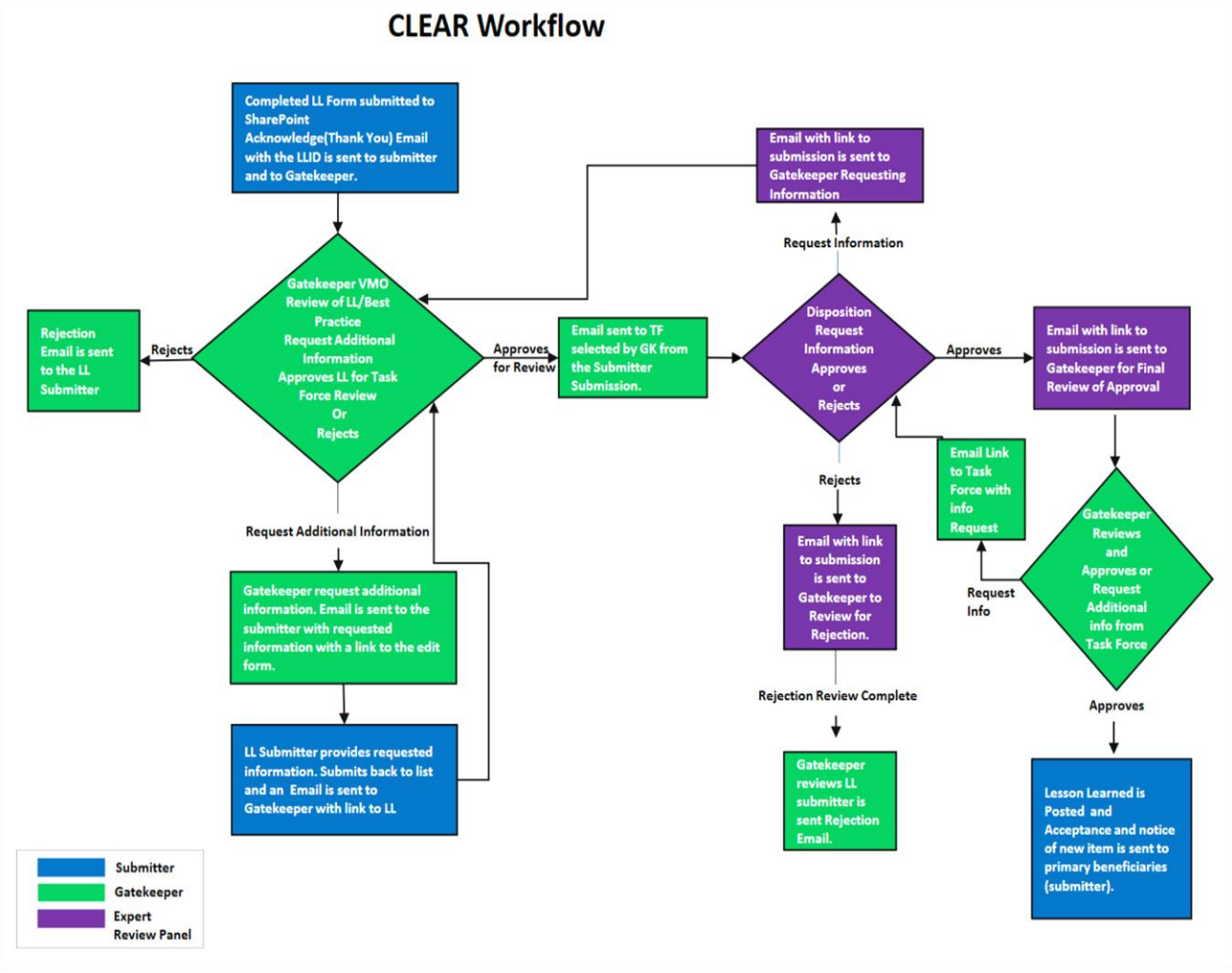

Figure 1. CLEAR workflow process.

Information Entry

End-users can submit knowledge gained from their routine work into an internal-only web-based CLEAR portal that was developed using a Design for Six Sigma approach on a Microsoft SharePoint portal, with Microsoft Access as the backend for storing and retrieving data (Banerjee,

Jaselskis, & Alsharef, Design for six sigma (DFSS) approach for creating clear lessons learned database, 2020). To store the knowledge gained, end-users can use one or more of three submission forms that best suit(s) the information being entered: lessons learned, best practices or ideas, and/or solutions needed. All three forms require information to be entered as text in the English language. On the lessons learned form, the user inputs a description of the challenge or problem and the solution (if any) that was applied to overcome the issue. On the best practice or idea form, the user describes a best practice or idea implemented at the project site, along with an example solution being adopted by other state DOTs or transportation agencies. End-users enter information on these two forms about their experiences working on projects, whereas the third form allows them to reach out to their colleagues and solicit information about an issue or obstacle faced in a project. That is, on the solution needed form, the end-user puts forth the problem, and anyone who has encountered and/or solved a similar problem can come forward with a potential solution.

These three forms act to provide effective communication and disseminate knowledge among NCDOT personnel who need information at an appropriate and specific time. Although the CLEAR portal has a search feature for end-users to seek relevant content based on keywords, the results are based on the presence of the word within the lessons learned and best practices. As an improvement to this process, the developed AI-based model's search functionality is more intuitive and displays a ranked list of results that contain semantically similar content that is closest to the keyword or search phrase entered. The next section elaborates on this work and discusses the differences between the current search functionality and ways that the developed CD-SAIL model will help users identify and select the most relevant automatically and intelligently suggested CLEAR content.

Figure 2 presents a sample best practices entry describing the problem and example solution in a raw text format. In addition, each CLEAR entry form has a provision to enter metadata, meaning that end-users can supplement information by adding relevant files such as PDFs, images, and email attachments. The metadata helps the end-users peruse knowledge in order to yield additional relevant information. However, for the scope of this paper, the results are limited to the raw text, although analyzing and automating information dissemination using metadata is a new dimension to explore in the future.

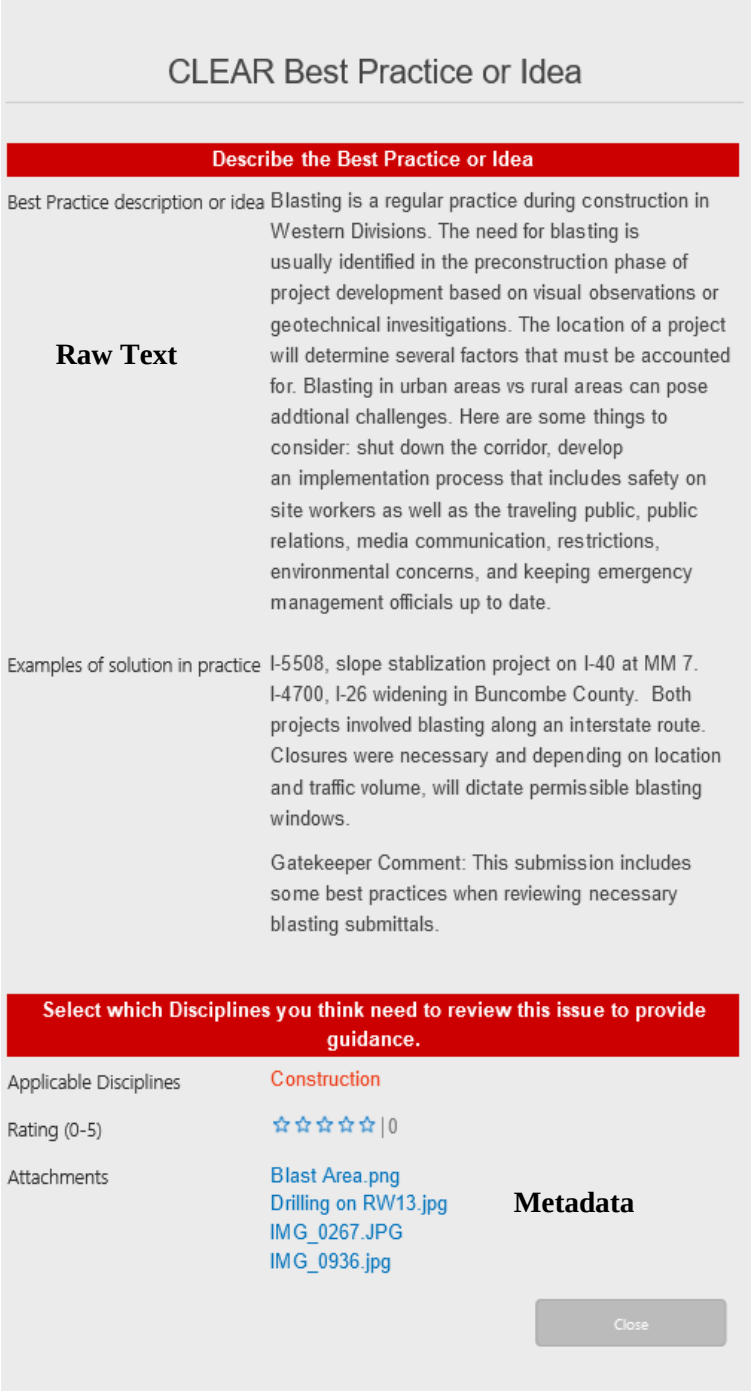

Figure 2. Example of best practice entry within CLEAR.

Institutionalizing knowledge and internal innovation

At the outset of the CLEAR database entry process, the end-user initiates a knowledge entry either in the form of a lesson learned or best practice. The gatekeeper verifies the completeness of the

entry and passes it along to the Expert Review Panel for its review and vetting before sending its disposition to the gatekeeper. The gatekeeper then uploads the entry into the CLEAR portal. The time period for a submission to be uploaded as an accepted entry and appear in the CLEAR portal generally is ninety days from the time it is first submitted by the end-user. Once the entry has been uploaded into the portal, it is available for other users to peruse and apply to their projects, thus paving the way for enhanced communication and knowledge-sharing.

In the next steps, the gatekeeper, the Technical Advisory Group, and the innovation coordinators periodically review new CLEAR entries by deliberating each entry's potential organizational innovation to bring about changes in the existing workflow processes. The selected entries with a potential to innovate are then shared among project teams and units by the respective innovation coordinators for the widespread adoption of the accepted lesson learned and/or best practice and to maximize outreach efforts. Finally, these entries are flagged as institutional knowledge with the potential to spur internal organizational innovation, thereby helping the NCDOT to retain its market competitiveness. Note, however, that not all entries within the CLEAR program are geared towards internal innovation, but the intent is to maximize such entries for the long-term benefit of the NCDOT. Figure 3 shows the various steps involved for an entry to be converted to institutional knowledge, possibly leading to positive changes in the workflow processes.

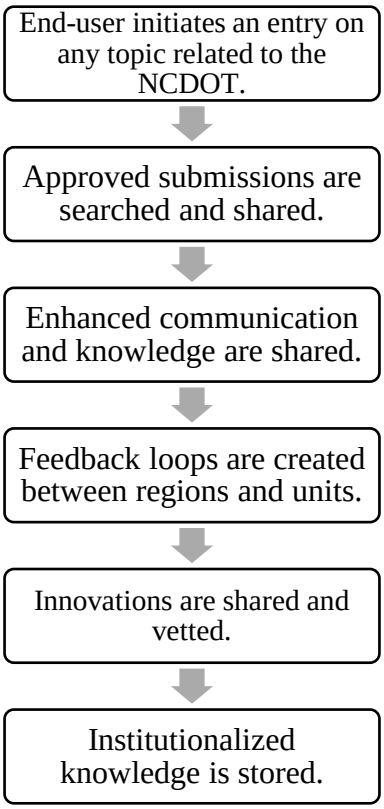

Figure 3. Steps involved in a CLEAR entry for institutionalizing knowledge.

4. ARTIFICIAL INTELLIGENCE (AI) MODEL

A successful lessons learned program is one where end-users are able to make effective use of the knowledge stored within the repositories for future projects. Numerous databases have become defunct for want of end-users to embrace and use them. Dalton (Dalton, 2013) states that organizations find it increasingly difficult for end-users to look into and extract knowledge from these lessons learned databases and analogizes the situation as a black hole where information is lost forever, rendering all previous efforts useless and risking repeat mistakes over extended periods of time, thus causing financial loss. The dominant mode for knowledge extraction from lessons learned databases is a keyword-based search. This method requires exact words to be specified in the lessons learned narrative for extraction. The choice of keywords is up to the user and determines the quality of relevant retrieved lessons learned. Lessons learned do not directly incorporate the entire context of the project, which includes a variety of factors (location, environment, materials, timeline, resources, etc.). This research used the latest advances in computational language models to address this challenge.

Machine learning systems, especially those used for natural language processing, require extensive input data to train effective models. The research team at NCSU compiled a comprehensive collection of transportation construction texts from various sources. This dataset includes over 4,000 documents and more than 1.5 million words. The texts were sourced from the CLEAR database (including both lessons learned and best practices), a sample of eight NCDOT projects, the NCDOT construction manual, several textbooks, and thousands of claims and change orders related to NCDOT projects. This approach enabled the NCSU research team to train a Doc2Vec model (Le & Mikolov, 2014). This statistical language model gained a fundamental understanding of the intricacies of transportation construction texts by learning the statistical co-occurrence of terms from the training corpus relevant to this specific domain.

We used Google's Tesseract optical character recognition software (Smith R. , 2007) in order to parse text out of PDFs that had been scanned or otherwise missing the markup needed to extract text directly. The team then used a custom tokenization run-time to parse the results into a stream of tokens (words) per document. For other sources, the researchers used a combination of loading directly from the CLEAR database and scripts to crawl manuals from NCDOT web

pages. Figure 41 shows the methodology adopted to suggest the ranked list of lessons learned/best practices (LL/BP in the figure) from the corpus of NCDOT project documents.

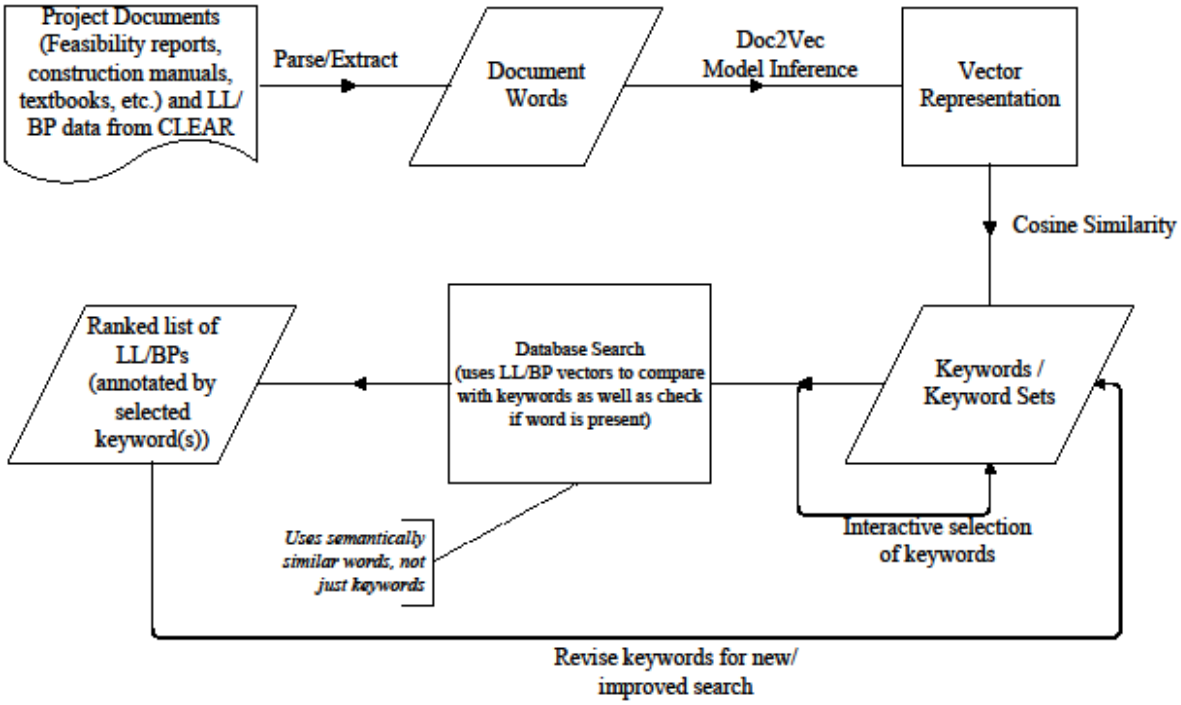

Figure 4. Steps involved in the creation of artificial intelligence model.

Doc2Vec simultaneously trains continuous vector representations of documents and words. The resulting vector space has useful semantic properties, such as a measure of similarity that uses cosine distance. In addition, vector addition and subtraction yield intuitive semantics. One of the canonical examples is *Paris* is to *France* as *Berlin* is to *Germany*, which can be computed using vector arithmetic.

$$V_{Paris} - V_{France} + V_{Germany} \approx V_{Berlin}$$

The developed artificial intelligence (AI) model contains transportation construction-specific information, such as *Power* is similar to *Transmission*, *Copper*, *Storing*, and *Energy*. Using the vector addition, *Power* and *Overhead* combine to be similar to *Powerline* and *Transmission*.

$V_{Power} + V_{Overhead} \approx V_{Powerline}$ or $V_{Transmission}$

Using subtraction shows that the sense of *Transmission* without the context of *Power* is more similar to *Axle*.

$$V_{Transmission} - V_{Power} \approx V_{Axle}$$

These examples help illustrate that the AI model understands the different contexts of common words found in transportation construction text. This feature is one of the advantages of curating a corpus instead of using a more generic pre-trained model. Other semantic examples would be the model's ability to understand that *Interchanges* and *Intersections* are related and that *Fiberoptics* is related to *Roadways*.

The NCSU research team tagged documents in the Doc2Vec model both uniquely and by their source. This effort led to vector representations for a project and the individual documents that the project contains. For example, a project's feasibility study can be distinguished from its environmental impacts.

As the goal of this project is to facilitate access to lessons learned and best practices, the NCSU research team utilized the developed AI model to facilitate the process of identifying which lessons learned and best practices are relevant to a project. The general flow of the system is to upload the documents (primarily the feasibility study) for a project. Next, the system extracts text from the documents and uses the language model to generate vector representations for each document and the overall project. This representation is used to generate a set of keywords based on cosine similarity. Each keyword corresponds to a larger set of similar words, such as *Utility*, *Utilities*, etc., which are specifically tagged if they are present in a document, or merely inferred as being related. A prime example would be the model understanding the semantic similarity between *Interchanges* and *Intersections*. This ability allows far larger relevant sets to be detected than a manual keyword selection approach, or an approach that only returns keywords that are present in a document. This process is interactive where the user (typically a project manager or other staff) can add or remove keywords based on the user's understanding of the project and goals for using the system. Table 5 provides word similarities that can be used for keyword sets. Note that both the keywords and notable similar words exclude morphological variants, which the model also marks as similar.

Table 1. Word Similarities Used for Keyword Sets

Keywords	Notable Similar Words from Model
Resurfacing	Grading, Widening, Pavement, Reclamation, Undercut
Power, Powerlines	Transmission, Copper, Energy, Storing, Electricity, Overhead
Underground	Leaking, Tunnels, Powerlines
Water	Sewer, Agitator, Discharges
Water, Sewer	Leaking
Soil, Contamination	Unstabilized, Toxic, Siltation, Hazardous
Widening	Roundabouts, Interchange, Improvements, Resurfacing
Road, Roadway	Avenue, Rd, Roadbed, Vehicles
Fiber, Fiberoptic	Cable, Utilities, SCP (fiber technology), YAGI (brand of cable), filtering, roadways

Beyond its ability to identify similar words, the model is robust in finding common misspellings. Consider, for example, the word *Utility*. Documents may contain the misspellings *Utilitiy*, *Utiltiy*, and *Utlility*, but because these words are all used in the same context and manner as the correct spelling, *Utility*, the model correctly infers that these words are semantically the same.

In the final step, the system automatically searches the lessons learned database for the selected keyword sets. This step returns a ranked order of lessons learned and the respective relevant keywords. This automatic search allows the user to determine quickly which keywords are the most relevant, ideally making for a better user experience. The user is also free to revise selected keywords and see updated results. The database search utilizes the same language model that is used to generate the keywords, which allows the lessons learned to be matched to projects/keywords even if the specific word is not present. An example is a lesson learned about *Power* and *Powerlines* that potentially matches with the keyword *Utility* because these words are related terms in the language model, even if none of these words is present in either the project documents or a particular lesson learned. This increased flexibility greatly enhances the ability of

the AI model to make accurate recommendations without the writers of the lessons learned having to identify keywords or the users having to fine-tune keyword searches. These improvements and reductions in user burden will make CLEAR more intuitive and thus increase the likelihood that project teams will find pertinent information, thereby saving money and time for the NCDOT.

Our search methodology differs from the original CLEAR implementation and naive keyword search approach. In a naive keyword search, each entry in the CLEAR database is searched in order to find exact matches of the keywords input in the entry text. This process can be computationally efficient using modern database technologies but fails to find many relevant matches because of the requirement to find an exact match. The original algorithm used in the CLEAR program is approximate string-matching (sometimes referred to as a fuzzy search). This algorithm is a popular choice for web search interfaces that often yield better results than a naive search. This algorithm again searches each database entry for keyword matches, but this time allows partial matches. For instance, 'road' matches 'roads' (with a minor penalty) because it differs by only one letter. This ability to match partially is particularly useful for misspelled words and typos, but semantically dissimilar words with similar spellings can cause problems.

The search methodology in the CD-SAIL model first performs an exact keyword search and then compares the semantic meaning of each search term (and the overall search phrase) to the computed semantic meaning of each database entry. This semantic closeness is computed via the cosine similarity of the relevant word, document, and sentence vectors. Figure 6 shows the search results obtained from the CD-SAIL model. In this particular example, the user has searched for '*utilities*'. The ranked list of CLEAR database entries that pertain to the search phrase '*utilities*' is displayed for the user to peruse and explore further by clicking on the most relevant search result that is yielded. Additionally, the model suggests semantically similar keywords that are based on cosine similarity to the search term that is input by the user. Therefore, chances are improved that the end-user can fine-tune the search requirements to acquire the specific information they seek from the CLEAR database to be applied to future projects.

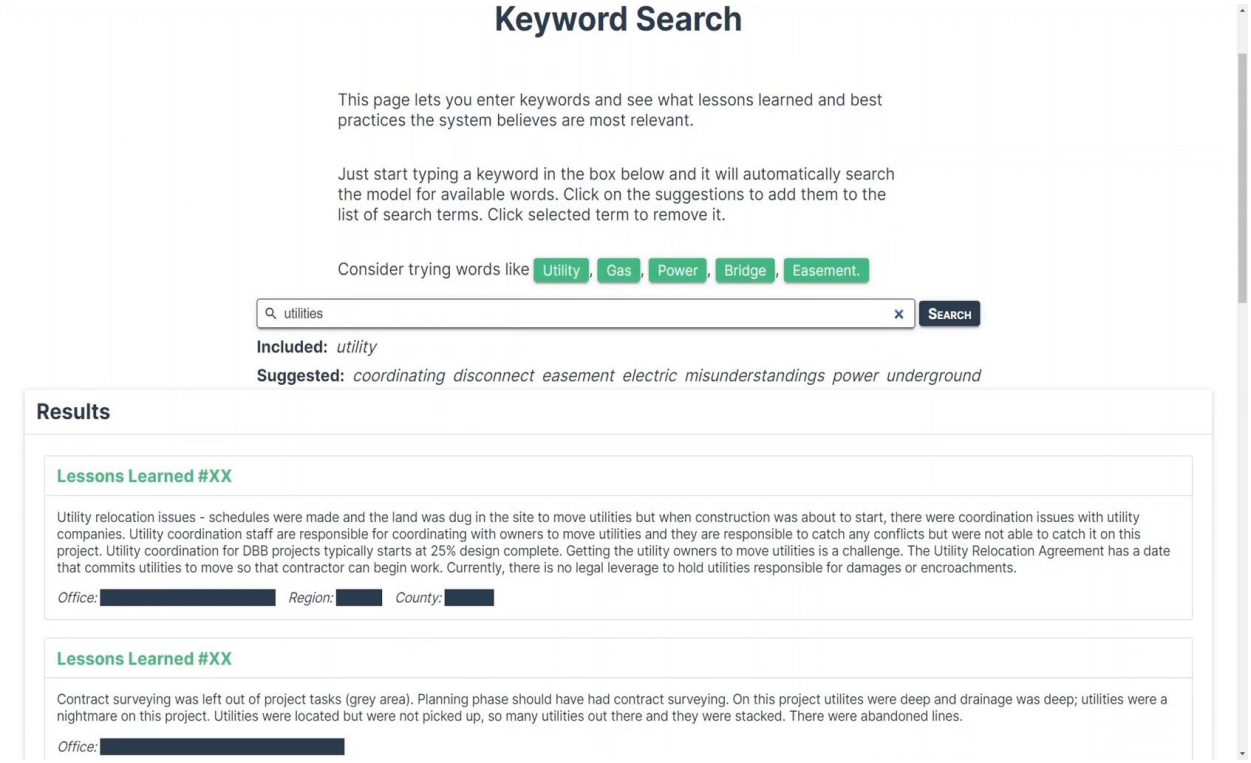

Figure 5. Search results from CLEAR database based on keyword search input by the end-user.

5. VALIDATION

We evaluated the CD-SAIL model for the CLEAR database in several ways. As shown in Table 1, we used construction domain-specific knowledge extensively to sample the language model and validate the results. We repeatedly chose commonly used transportation construction terms and checked similar words and documents from the trained model. We iterated through many different models to arrive at one that accurately reflects the expert domain knowledge obtained from the training corpus. Next, we evaluated the use of this model to search the CLEAR database of lessons learned and best practices by having experts from the NCDOT and our team test the model. Each user systematically entered search queries and evaluated the results for relevance. This process was repeated twice, once using the existing and currently deployed approximate string-matching algorithm and again using our AI-assisted search algorithm in the CD-SAIL model. The model successfully ranked relevant lessons learned and best practices higher in the results than the approximate string-matching algorithm and found entries that were not returned by the approximate string-matching algorithm. The reason for the benefits of our model is twofold. First, approximate string-matching can inadvertently assume that ‘roadway’ and

‘railway’ share the most letters and thus are related, whereas our model correctly understands the nuanced meanings of ‘roadway’, ‘railway’, and ‘railroad’. Furthermore, our model can successfully navigate many more such examples of common search terms. After our evaluation and validation of the CD-SAIL model, the NCDOT has expressed interest in formally deploying this model for its users. We are in the midst of this formal adoption process. The reader is encouraged to note that the model’s name, CD-SAIL, is used for the purposes of this paper and that, as more information becomes available, the model’s current name may be revised in future.

Significance of the Artificial Intelligence Model Applied to CLEAR

The research team used both the trained model and a more traditional string search algorithm for searching the lessons learned database. The researchers reviewed the search terms entered and first scanned the entire collection of lessons learned/best practices for exact word and phrase matches. Next, they loosened the criteria and searched for substring matches. These cases are important to distinguish, particularly in a technical language context, because, for example, an exact match for *Road* is significant but a substring match for *Road* in *Railroad* would yield incorrect results. At this stage, the language model begins to play a role. After prioritizing exact matches to match the expected behavior for users, the researchers used the language model to look at the sense of each word individually, which allowed the distinction to be made between *Road* matching *Roadway* (relevant) and *Road* matching *Railroad* (not relevant). The research team also took advantage of the Doc2Vec model’s ability to model documents to compare the semantic meaning of the input search phrase with the semantic meaning of the lessons learned/best practices. For example, consider that *Wet Utilities* has a more specific meaning than *Utilities* by itself. In this example, matches for *Utilities* only would not correspond to the semantic intent of a search for *Wet Utilities*. In the end, a final list of results can be derived from the combined ranking of exact, partial, and language model scores (Banerjee S. , Potts, Jhala, & Jaselskis, 2021).

Usability Study with NC DOT Project Managers

We conducted a usability survey to validate the efficacy and usefulness of the cognitive search functionality for the CD-SAIL model. Figure 6 lays out the study protocol that was used for the surveys. Recruitment was done through careful anonymized selection from recruitment emails

sent out by the research team to former project managers familiar with projects that had documented LL/BPs in the database.

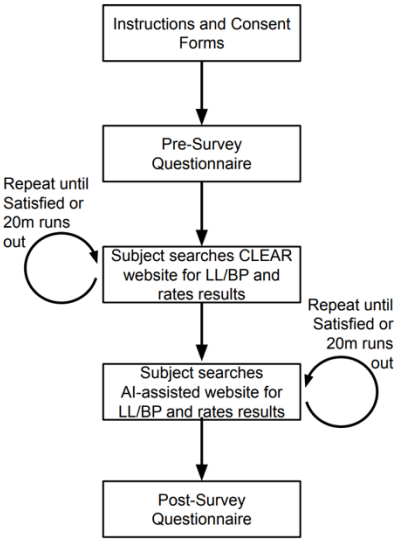

Figure 6. Study Protocol for the Usability Survey

The first participants were given a consent form and given details and the purpose of the study. This was followed by a pre-survey questionnaire to ask them questions about their familiarity, experience, and process for LL/BP documentation as a baseline. They were given specific projects and questions related to these projects that would necessitate a search in the CLEAR database. They were first only given the CLEAR site and used keyword-based search. They provided ratings for relevance and quality of the top search results. Then they repeated the process with the AI-assisted interface (Figure 5) and when satisfied with their results, were given a post-survey questionnaire. The search steps were capped at 20 minutes.

Overall, the results showed that among all raters, 76% of the top 5 results from the AI-assisted search model were rated as highly relevant against 44% of the results from the keyword-based search. We also evaluated per search term across models. While CD-SAIL performed overall much better (blue-right side in Figure 7), there were terms like *Matting* that it did not do well on due to a lack of sufficient references in the relevant documents. In this case, there was only one result so the 4 other results in the top 5 were not relevant.

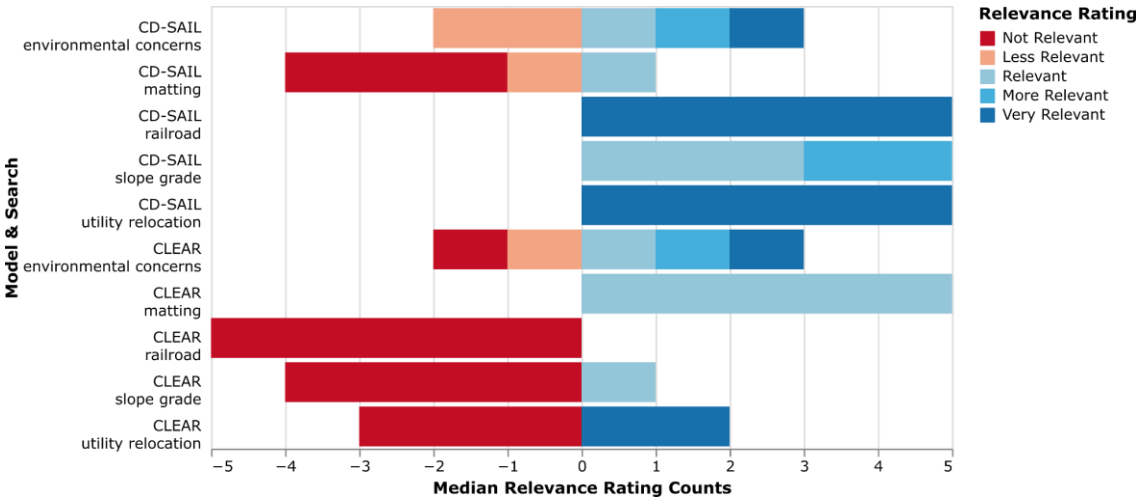

Figure 7. Median relevance ratings across different topics related to projects in the study.

6. DISCUSSION

Well-managed knowledge management practices help to retain valuable information that can help minimize repeated mistakes when applied to future projects. Lessons learned databases have been in existence for the past two decades to help organizations store and retrieve knowledge. However, most of these databases have been rendered ineffective or obsolete for want of being embraced by end-users. End-users must be encouraged to contribute their knowledge and make use of stored knowledge repositories to apply to projects because failure to do so risks having a futile knowledge management program and poor institutional knowledge.

The developed CD-SAIL model automatically and intelligently suggests lessons learned and best practices that are stored within the CLEAR database at the NCDOT. As far as transportation infrastructure projects are concerned, better communication across stakeholders must be facilitated by harnessing the extant information that is found in domain-specific corpora. Looking beyond the specific goals of the CLEAR program, this work contributes to advances in computing technology and impactful knowledge-intensive applications. It also fills the knowledge gap by facilitating deeper human-to-computer and computer-mediated human-to-human communication in the realm of creating and maintaining robust knowledge repositories.

Steps and considerations for deploying the CD-SAIL model in an organizational setting.

The first step in deploying the CD-SAIL model is to identify all the major sources of text that relate to practices within the NCDOT, such as project feasibility study reports, construction manuals, project contract documents, and domain-specific textbooks. Although most of these sources are machine-readable PDFs, a few contract documents are much older and type-written, which makes it difficult for a computer to decipher the language directly. We used Google Tesseract, a popular open-source optical character recognition engine, to enable the computer to read and extract text from such documents. We prepared a corpus of more than 1.5 million unique document words that then were represented in high-dimensional abstract vector space using a Doc2Vec model. These words were clustered based on their semantic similarity via cosine distance. Creating such clusters of words helps the model to identify CLEAR database entries based on a word being close in meaning to the searched keyword(s), even though the exact keyword does not appear in the entry. For instance, the model is able to recognize that ‘utilities’ and ‘misunderstandings’ are probabilistically semantically related as they appear in raw textual

sources, although such an association would not have been possible using a general word corpus in the English language. The neural network language model described in this paper is robust in dealing with context-specific semantically similar words, thereby strengthening the search results compared to an objective domain ontology search approach (Rezgui, 2006) or rule-based reasoning using natural language processing that is based on domain knowledge (Zhang & El-Gohary, 2016). Moreover, using advanced probabilistic models and providing a domain-specific word corpus eliminates the need to predetermine the relationships between words to build ontologies manually (Kim & Chi, 2019) and reduces the need for computational resources by using dimensional reduction techniques that form conditional associations between words that are present in the text source (Bengio, Ducharme, Vincent, & Jauvin, 2003). In short, eliminating the human effort needed to create semantic associations will yield improved and reliable search results that are beneficial to end-users.

The developed CD-SAIL model has been validated by NCDOT end-users and fine-tuned based on inputs received during the validation stage. The validation results show that the CD-SAIL model is able to accurately reflect the ranked list of CLEAR database entries based on the searched keyword that is input by the end-user, thereby making it easy to look for the most relevant content and apply the knowledge to future projects. Considering the fact that personnel are strained for time, given their routine work schedules, the CD-SAIL model is expected to reduce the burden on end-users by automatically suggesting the most relevant documents and sources so that users can peruse the stored knowledge quickly and efficiently. By doing so, this AI model is expected to keep end-users engaged with the CLEAR program, thereby maximizing the chances of the success of the program in the long run. Importantly, the NCDOT will benefit as an organization by spurring internal innovation, leading to enhanced institutional knowledge and workflow processes.

A worthwhile future endeavor would be for other DOTs that may not have functional knowledge repositories to create such databases and apply the developed AI model to intelligent and automated knowledge suggestion systems. Such applications would be beneficial to both the end-users in terms of being able to peruse the most relevant content and to DOTs in terms of increasing internal innovation through the sharing and application of knowledge gleaned by the end-users in their routine work.

7. CONCLUSIONS

Although organizations have started to realize the benefits of having effective knowledge repositories in place, such repositories bear a huge risk of failure if they are not accepted by end-users. Once end-users start neglecting to use these knowledge storage and retrieval mechanisms, the repository will become defunct. Therefore, knowledge repository designs must incorporate the fact that time and quality are of utmost importance as they pertain to the end-users' desire to peruse the knowledge within the repository. That is, the end-user must be able to receive the most relevant knowledge they are seeking within a relatively short period of time. As such, knowledge repositories must become more user-friendly to maximize their reach and minimize the chance of program failure.

This project developed a novel effort to use the latest AI tools to automate the process of intelligently disseminating knowledge through a neural network language model to benefit the recently created CLEAR program for the NCDOT. The Construction Domain-Specific Artificial Intelligence (CD-SAIL) model was trained using a domain-specific corpus of words that were extracted from several sources of relevant texts, including CLEAR entries, NCDOT construction specifications and manuals, contract documents, and construction textbooks. Within the transportation and public infrastructure domain, creating a domain-specific word corpus is even more important due to the scope and societal impacts of capital projects.

The developed neural network language model, identified in this paper as the CD-SAIL model, can probabilistically map essential keywords/sets to text documents and suggest the most relevant and necessary documents that are semantically related to such automatically detected keywords in these project documents. The benefits are two-fold. First, the developed model will help end-users sift through large volumes of data quickly to peruse only the most pertinent data, thereby saving their time and energy. Second, it will encourage end-users to use the CLEAR database more often, thereby minimizing the risk of program failure due to a lack of end-user participation. The CD-SAIL model was validated by project managers of various NCDOT projects to obtain comprehensive feedback. In response, we made the necessary modifications to fine-tune the model. Ultimately, applying AI using natural language processing complements the accurate analysis of text content within the CLEAR database. In the long run, through its CLEAR program, the NCDOT will benefit from the organizational innovation that arises out of well-maintained institutional knowledge.

8. REFERENCES

AASHTO Journal. (2021, May 14). *State DOT panel examines workforce recruitment, retention issues*. Retrieved December 18, 2021, from https://aashtojournal.org/2021/05/14/state-dot-panel-examines-workforce-recruitment-retention-issues/

Alsharef, A. F. (2015). *Design of a Construction Expenditure Forecasting and Monitoring Tool for NCDOT Mega Projects.* Raleigh: North Carolina State University.

Alsharef, A., Banerjee, S., Uddin, S. M., Albert, A., & Jaselskis, E. (2021). Early impacts of the COVID-19 pandemic on the united states construction industry. *International Journal of Environmental Research and Public Health, 18*(4), 1559. doi:10.3390/ijerph18041559

Alsharef, A., Banerjee, S., Uddin, S. M., Albert, A., & Jaselskis, E. (2021). Early Impacts of the COVID-19 Pandemic on the United States Construction Industry. *International Journal of Environmental Research and Public Health, 18*(4), 1559. doi:https://doi.org/10.3390/ijerph18041559

Amir, R., & Parvar, J. (2014, February). Harnessing knowledge management to improve organizational performance. *International Journal of Trade, Economics and Finance, 5*(1), 31-38. doi:10.7763/IJTEF.2014.V5.336

Anderson, S. D., & Tucker, R. L. (1994). Improving Project Management Of Design. *Journal of Management in Engineering, 10*(4), 35-44.

Anumba, C., Egbu, C., & Carrillo, P. (2005). *Knowledge management in construction.* Oxford: Blackwell.

Assaad, R., & El-adaway, I. H. (2021). Guidelines for responding to COVID-19 pandemic: best practices, impacts, and future research directions. *Journal of Management in Engineering, 37*(3), 06021001. doi:10.1061/(ASCE)ME.1943-5479.0000906

Banerjee, S., Jaselskis, E. J., & Alsharef, A. (2020). Design for six sigma (DFSS) approach for creating clear lessons learned database. *Periodica Polytechnica Architecture, 51*(1), 75-82. doi:https://doi.org/10.3311/PPar.15442

Banerjee, S., Jaselskis, E. J., & Alsharef, A. F. (2020). Design For Six Sigma (DFSS) Approach for Creating CLEAR Lessons Learned Database. *Periodica Polytechnica Architecture, 51*(1), 75-82. doi:https://doi.org/10.3311/PPar.15442

Banerjee, S., Potts, C. M., Jhala, A. H., & Jaselskis, E. J. (In press). Neural language model based intelligent semantic information retrieval on NCDOT projects for knowledge management. *Proceedings of the 2021 ASCE International Conference on Computing in Civil Engineering (i3CE2021).* Orlando, FL: American Society of Civil Engineers.

Banerjee, S., Potts, C., Jhala, A. H., & Jaselskis, E. J. (2021). Neural Language Model based Intelligent Semantic Information Retrieval on NCDOT Projects for Knowledge Management. *International Conference on Computing in Civil Engineering (i3CE2021)* (p. forthcoming). Orlando: American Society of Civil Engineers.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137-1155.

Berka, P. (2011). NEST: A compositional approach to rule-based and case-based reasoning. *Advances in Artificial Intelligence, 2011*, 1-15. doi:10.1155/2011/374250

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146. doi:10.1162/tacl_a_00051

Carrillo, P., & Anumba, C. (2002). Knowledge Management in the AEC Sector: An Exploration of the Mergers and Acquisitions Context. *Knowledge and Process Management, 9*(3), 149-161.

Clark, K., & Hammer, M. (2008, December). Communities Of Practice:The VDOT Experience. *Knowledge Management Review, 11*(5), 10-15. Retrieved October 26, 2020

Colorado Department of Transportation. (2018). *Lean Everyday Ideas*. Retrieved May 25, 2020, from https://www.codot.gov/business/process-improvement/lean-everyday-ideas

Construction Industry Institute. (2017). *CII Best Practices Handbook* (Vols. SP166-4). Austin, TX: Construction Industry Institute.

CROSS-US. (2020, April). *Structural Safety::Confidential Reporting on Structural Safety*. Retrieved May 25, 2020, from https://www.cross-us.org/about-us/

Dalton, J. (2013, October 17). *Can CMMI Save Us from the Black Hole of Lessons Learned?* Retrieved from http://askthecmmiappraiser.blogspot.com/2013/10/can-cmmi-save-us-from-black-hole-of.html

Dekker, R., & de Hoog, R. (2000). The monetary value of knowledge assets: a micro approach. *Expert Systems with Applications, 18*, 111-124.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the NAACL-HLT 2019* (pp. 4171-4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Egbu, C. (2004). Managing knowledge and intellectual capital for improved organizational innovations in the construction industry: an examination of critical success factors. *Engineering, Construction and Architectural Management, 11*(5), 301-315. doi:10.1108/09699980410558494

El-Rayes, K., Liu, L., El-Gohary, N., Golparvar-Fard, M., & Ignacio, J. E. (2017). *Best Management Practices And Incentives To Expedite Utility Relocation.* Chicago: Illinois Center for Transportation. doi:https://doi.org/10.36501/0197-9191/17-017

Everett, J. G., & Farghal, S. (1994). Learning curve predictors for construction field operations. *Journal of Construction Engineering and Management, 120*(3), 603-616. doi:https://doi.org/10.1061/(ASCE)0733-9364(1994)120:3(603)

Federal Highway Administration. (2018, September 04). *Summary of Lessons Learned from Recent Major Projects*. Retrieved May 20, 2020, from https://www.fhwa.dot.gov/majorprojects/lessons_learned/lessons_learned.cfm

Ferrada, X., Nunez, D., Neyem, A., Serpell, A., & Sepulveda, M. (2015). A lessons-learned system for construction project management: a preliminary application. *Proceedings of the 29th World*

Congress International Project Management Association (IPMA). 226, pp. 302-309. Westin Playa Bonita, Panama: Procedia - Social and Behavioral Sciences. doi: 10.1016/j.sbspro.2016.06.192

Fong, P. S., & Yip, J. C. (2006). An Investigative Study of the Application of Lessons Learned Systems in Construction Projects. *Journal for Education in the Built Environment, 1*(2), 27-38.

Fullerton, C. E. (2020, February 7). *Summary report of the progress from 2019 and upcoming goals for 2020*. Retrieved April 08, 2020, from CLEAR Program Report: https://connect.ncdot.gov/projects/Value-Management/CLEAR-Program/Documents/CLEAR%20Program%20Report%20Feb%202020.pdf

Fullerton, C. E., Tamer, A. W., Banerjee, S., Alsharef, A. F., & Jaselskis, E. J. (2021). Development of North Carolina Department of Transportation's CLEAR Program for Enhanced Project Performance. *Transportation Research Record*. doi:https://doi.org/10.1177/0361198121995195

Fullerton, C. E., Tamer, A. W., Banerjee, S., Alsharef, A., & Jaselskis, E. J. (2021). Development of north carolina department of transportation's clear program for enhanced project management. *Transportation Research Record, 2675*(7), 222-234. doi:https://doi.org/10.1177/0361198121995195

Ganopol, A., Oglietti, M., Ambrosino, A., Patt, F., Scott, A., Hong, L., & Feldman, G. (2017, September). Lessons Learned: An Effective Approach to Avoid Repeating the Same Old Mistakes. *Journal of Aerospace Information Systems, 14*(9). doi:10.2514/1.I010485

Gibson Jr., G., Caldas, C., Yohe, A., & Weerasooriya, R. (2008). *An Analysis of Lessons Learned Programs in the Construction Industry.* CII Research Report.

Goh, S. (2002). Managing effective knowledge transfer: an integrative framework and some practice implications. *Journal of Knowledge Management, 6*, 23-30. doi:10.1108/13673270210417664

Goodrum, P. M., Yasin, M. F., & Hancher, D. E. (2003). *Lessons Learned System for Kentucky Transportation Projects.* Kentucky Transportation Center Research Report.

Goodrum, P., Smith, A., Slaughter, B., & Kari, F. (2008). Case Study and Statistical Analysis of Utility Conflicts on Construction Roadway Projects and Best Practices in Their Avoidance. *Journal of Urban Planning and Development, 134*(2), 63-70. doi:https://doi.org/10.1061/(ASCE)0733-9488(2008)134:2(63)

Grant, D., & Mergen, E. A. (2009). Towards the use of Six Sigma in software development. *Total Quality Management, 20*(7), 705-712.

Hansen, M. T., Nohria, N., & Tierney, T. J. (1999, March). *What's Your Strategy for Managing Knowledge?* Retrieved February 23, 2020, from Harvard Business Review: https://hbr.org/1999/03/whats-your-strategy-for-managing-knowledge

Henard, D. H. (2020). *CLEAR (Communicate Lessons, Exchange Advice, Record) Technology Transfer and Metric Building.* Research and Development Unit. Raleigh: North Carolina Department of Transportation.

Ho, S.-P., Tserng, H.-P., & Jan, S.-H. (2013). Enhancing Knowledge Sharing Management Using BIM Technology in Construction. *The Scientific World Journal, 2013*, 1-10. doi:10.1155/2013/170498

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: industrial-strength natural language processing in python. Forthcoming. doi:10.5281/zenodo.1212303

Hu, M., Pieprzak, J. M., & Glowa, J. (2004). Essentials of Design Robustness in Design for SixSigma (DFSS) Methodology. *SAE 2004 World Congress & Exhibition*, (p. 13). doi:https://doi.org/10.4271/2004-01-0813

International Atomic Energy Agency. (2011). *Design Lessons Drawn From The Decommissioning Of Nuclear Facilities.* Vienna: IAEA Publishing Section. Retrieved May 20, 2020, from https://www-pub.iaea.org/MTCD/Publications/PDF/TE_1657_web.pdf

Issa, R. R., & Haddad, J. (2008). Perceptions of the impacts of organizational culture and information technology on knowledge sharing in construction. *Construction Innovation, 8*(3), 182-201. doi:10.1108/14714170810888958

ITS Joint Program Office. (2020, May 27). *Lessons Learned Overview*. Retrieved May 29, 2020, from https://www.itslessons.its.dot.gov/its/benecost.nsf/LessonHome

Johari, S., & Jha, K. N. (2021). Learning curve models for construction workers. *Journal of Management in Engineering, 37*(5), 04021042. doi:10.1061/(ASCE)ME.1943-5479.0000941

Kim, T., & Chi, S. (2019). Accident case retrieval and analyses: using natural language processing in the construction industry. *Journal of Construction Engineering and Management, 145*(3), 04019004. doi:10.1061/(ASCE)CO.1943-7862.0001625

Knoco. (2009, May 2009). *The status of lessons learning in organisations.* Retrieved April 28, 2018, from Knoco White Paper - Lessons Learned Survey: https://www.knoco.com/Knoco%20White%20Paper%20-%20Lessons%20Learned%20survey.pdf

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning. 32*, pp. 1188-1196. Beijing, China: JMLR: W & CP.

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning. 32*, pp. 1188-1196. Beijing, China: JMLR: W & CP.

Lee, J. H., Yi, J. S., & Son, J. W. (2019). Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based nlp. *Journal of Computing in Civil Engineering, 33*(3), 04019003. doi:10.1061/(ASCE)CP.1943-5487.0000807

Li, X., Shen, G. Q., Wu, P., & Yue, T. (2019). Integrating Building Information Modeling and Prefabrication Housing Production. *Automation in Construction, 100*, 46-60. doi:10.1016/j.autcon.2018.12.024

Lin, J. J., & Golparvar-Fard, M. (2020). Construction Progress Monitoring Using Cyber-Physical Systems. In A. Chimay, & N. Roofigari-Esfahan, *Cyber-Physical Systems in the Built Environment* (pp. 63-87). Champaign: Springer. doi:10.1007/978-3-030-41560-0_5

McCullouch, B. G., & Patty, R. (1994). *An INDOT Lessons Learned Cconstructability Program And Integradted Multimedia System.* Final report, Purdue University.

Mehrbod, S., Staub-French, S., Mahyar, N., & Tory, M. (2019). Characterizing Interactions with BIM tools and Artifacts in Building Design Coordination Meetings. *Automation in Construction, 98*, 195-213. doi:10.1016/j.autcon.2018.10.025

National Skills Coalition. (2021, May 25). *Building a people-centered infrastructure plan for the 21st century.* Retrieved January 10, 2022, from https://nationalskillscoalition.org/wp-content/uploads/2021/01/11102020-Building-a-People-Centered-Infrastructure-Plan-Memo-Public.pdf

NCDOT. (2004). *North Carolina general assembly NCDOT project delivery study.* Raleigh, NC: Dye Management Group. Retrieved from http://worldcat.org/oclc/191886690/viewonline

NCDOT Survey. (2020). *Innovation Culture Index*. Retrieved April 10, 2021, from https://ncdot.tech/

Neuendorf, K. A., & Kumar, A. (2015). Content Analysis. In G. Mazzoleni, *The International Encyclopedia of Political Communication* (pp. 1-10). John Wiley & Sons, Inc. doi:10.1002/9781118541555.wbiepc065

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37. Retrieved April 2, 2021, from http://www.jstor.org/stable/2635068

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543). Doha, Qatar: Association for Computational Linguistics.

Plotch, P. M. (2015). What's Taking So Long? Identifying the Underlying Causes of Delays in Planning Transportation Megaprojects in the United States. *Journal of Planning Literature, 30*(3), 282-295.

Potts, C. M., & Jhala, A. H. (2021). Narraport: narrative-based interactions and report generation with large datasets. *Proceedings of the International Conference on Interactive Digital Storytelling. 13138*, pp. 118-127. Tallinn, Estonia: Springer, Cham. doi:10.1007/978-3-030-92300-6_11

Project Management Institute. (2004). *A guide to the project management body of knowledge (PMBOK guide)* (6th ed.). Newton Square, Pa: Project Management Institute.

Quiroga, C., Kraus, E., & Overman, J. (2011). Strategies to Address Utility Challengesin Project Development. *Transportation Research Record: Journal of the Transportation Research Board, 2262*, 227-235. doi:10.3141/2262-23

Quiroga, C., McCleve, J., Lee, R., Kraus, E., Anspach, J., Sturgill, R., . . . Cooper, J. (2019). Strategic Research Needs in the Area of Utilities. *Centennial Papers*, 1-16. Retrieved from http://onlinepubs.trb.org/onlinepubs/centennial/papers/AFB70-Final.pdf

Rasoulkhani, K., Brannen, L., Zhu, J., Mostafavi, A., Jaselskis, E., Ryan, S., . . . Chowdhury, S. (2020). Establishing a Future-Proofing Framework for Infrastructure Projects to Proactively Adapt to

Complex Regulatory Landscapes. *Journal of Management in Engineering, 36*(4), 1-10. doi:https://doi.org/10.1061/(ASCE)ME.1943-5479.0000794

Rezaei, F., Khalilzadeh, M., & Soleimani, P. (2021). Factors Affecting Knowledge Management and Its Effect on Organizational Performance: Mediating the Role of Human Capital. *Advances in Human-Computer Interaction, 2021*, 1-16. doi:10.1155/2021/8857572

Rezgui, Y. (2006). Ontology-centered knowledge management using information retrieval techniques. *Journal of Computing in Civil Engineering, 20*(4), 261-270.

Robertson, J., Harrison, B., & Jhala, A. H. (2020). Interactive summarization for data filtering and triage. *Proceedings of the Thirty-Third International FLAIRS Conference (FLAIRS-33)* (pp. 252-257). North Miami Beach, FL: Association for the Advancement of Artificial Intelligence.

Saldana, J. (2015). *The Coding Manual for Qualitative Researchers.* SAGE Publications Ltd.

Sheehan, T., Poole, D., Lyttle, I., & Egbu, C. (2005). Strategies and business case for knowledge management. In C. Anumba, C. Egbu, & P. Carrillo, *Knowledge management in construction* (pp. 50-64). Oxford: Blackwell.

Smith, M. K. (2003). *Michael Polanyi and tacit knowledge*. Retrieved April 1, 2021, from The Encyclopedia of Pedagogy and Informal Education: https://infed.org/mobi/michael-polanyi-and-tacit-knowledge/

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (pp. 629-633). Curitiba, Brazil: IEEE. doi:10.1109/ICDAR.2007.4376991

Stamatiadis, N., Goodrum, P., Shocklee, E., Sturgill, R., & Wang, C. (2013). *Tools for Applying Constructability Concepts to Project Development (Design).* University of Kentucky.

Tserng, P. H., Yen-Liang, S. Y., & Lee, M. H. (2010). The use of knowledge map model in construction industry. *Journal of Civil Engineering and Management, 16*(3), 332-344. doi:10.3846/jcem.2010.38

Wenger, É., McDermott, R. A., & Snyder, W. (2002). *Cultivating Communities of Practice: A Guide to Managing Knowledge.* Boston: Harvard Business School Press.

Youndt, M. A., & Snell, S. A. (2004). Human resource configurations, intellectual capital, and organizational performance. *Journal of Managerial Issues, 16*(3), 337-360. Retrieved from http://www.jstor.org/stable/40604485

Zhai, C. X. (2008). *Statistical language models for information retrieval.* San Rafael, California: Morgan & Claypool Publishers.

Zhang, J., & El-Gohary, N. M. (2016). Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking. *Journal of Computing in Civil Engineering, 30*(2), 04015014. doi:10.1061/(ASCE)CP.1943-5487.0000346

Zhang, R., & El-Gohary, N. M. (2021). A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking. *Automation in Construction, 132*, 103834. doi:10.1016/j.autcon.2021.103834

Zou, P. X. (2007). A Longitudinal Study of E-learning for Construction. *Journal for Education in the Built Environment, 2*(2), 61-84. doi:https://doi.org/10.11120/jebe.2007.02020061

9. LIST OF APPENDICES

Appendix A. NCSU Institutional Review Board Materials for Human Subjects Study

NORTH CAROLINA STATE UNIVERSITY

INSTITUTIONAL REVIEW BOARD FOR THE USE OF HUMAN SUBJECTS IN RESEARCH SUBMISSION FOR NEW STUDIES

Protocol Number 24104

Project Title

AI-assisted keyword extraction for the CLEAR project

IRB File Number:

Original Approval Date:

07/28/2021

Approval Period

07/28/2021 - 01/01/2100

Source of funding (provide name of funder not account number):

North Carolina Department of Transportation

NCSU Faculty point of contact for this protocol:NB: only this person has authority to submit the protocol.

Arnav Jhala: Computer Science

Does any investigator associated with this project have a significant financial interest in, or other conflict of interest involving, the sponsor of this project? (Answer No if this project is not sponsored)

No

Is this conflict managed with a written management plan, and is the management plan being properly followed?

No

Preliminary Review Determination

Category:

Exempt d.2, d.3

In lay language, briefly describe the purpose of the proposed research and why it is important. Provide a brief synopsis of the study including who is targeted to participate and the data collection methods employed (limit text to 1500 characters)

We intend to observe whether an AI-assisted website provides better quality results and yields more engagement from

subjects on a web-search task for lessons learned from construction projects.

To this end, we designed an AI-assisted website to compare against an existing NCDOT website containing best practices and lessons learned for NCDOT projects. Participants will interact with both websites and indicate the quality of search results they receive. In addition, they will fill out a short survey indicating their overall rating of the interface.

Does any member of the project team who is responsible for the design, recruitment, consent, implementation of intervention, interaction with participants, or those handling identifiable private information under this IRB protocol - or any members of their immediate family (defined as spouse, dependent children - have any Significant Financial Interest or other types of conflict of interest (as described in SOP 14.3.a) related to the protocol?

If the answer is "yes," please provide the name of the investigator(s) with the potential or actual conflict and confirm that the relationship has been fully disclosed in the investigators most recent COI disclosure filed with NC State or disclosed through the collaborative research process. If there is a COI management plan in place with NC State University, please upload it with this application to ensure the IRB protocol meets the expectations of the COI plan and the COI is properly considered in the IRB review process. If you are uncertain how to respond or have questions, please contact COI-NOI-Compliance@ncsu.edu.

This research qualifies for Exemption. Review NC State's Exemption Research SOP for studies that may qualify. If you want to apply for an Exemption, download the Exemption Request Form and complete it. To the eIRB, upload the completed Exemption Request Form, all instruments, and if applicable a Data Access and Security Plan and the edited Consent/Opt-Out forms modified to fit the study design. Only complete the "Title" and "Description" tabs in the eIRB, upload the aforementioned documentation, and submit the eIRB application. Do not complete any other tabs within the eIRB system.

1

Is this research being conducted by a student?

No

Is this research for a thesis/disseration/capstone?

No

Is this research for a dissertatiion?

No

Is this independent research?

No

Is this research for a course?

No

Do you currently intend to use the data for any purpose beyond the fulfillment of the class assignment?

No

Please explain

If so, please explain

If you anticipate additional NCSU-affiliated investigators (other than those listed on the Title tab) may be involved in this research, list them here indicating their name and department.

Dr. Edward Jaselskis, CCEE

Siddharth Banerjee, CCEE

Will the investigators be collaborating with researchers at any institutions or organizations outside of NC State?

No

List collaborating institutions and describe the nature of the collaboration. If researchers from both institutions are doing any of the following activities: recruitment, consent process, data collection or handling of identifiable information/specimens a reliance agreement may be appropriate. For more information, please contact irb-coordinator-admin@ncsu.edu

North Carolina Department of Transportation

What is NCSU's role in this research?

Principal Investigator

Describe funding flow, if any (e.g. subcontractors)

Is this international research?

No

Identify the countries involved in this research

An IRB equivalent review for local and cultural context may be necessary for this study. Can you recommend consultants with cultural expertise who may be willing to provide this review? Consultants may not be a part of the research team or have a stake in the research project. Provide email contact information for consultant(s). A local context review may lengthen the time it takes for your approval.

Adults 18 - 64 in the general population?

Yes

NCSU students, faculty or staff?

No

Adults age 65 and older?

No

Minors (under age 18--be sure to include provision for parental consent and/or child assent). If minors are included in your research, please read through the NC State University Regulation for your additional responsibilities. Following this regulation is a requirement of your affiliation with NC State.?

No

List ages or age range:

Could any of the children be "Wards of the State" (a child whose welfare is the responsibility of the state or other agency, institution, or entity)?

No

Please explain:

Does this study involve people who are also incarcerated, involuntarily detained or committed, or are in a program or hospital as an alternative form of sentencing?

No

Pregnant women?

No

Are pregnant women the primary population or focus for this research?

No

Provide rationale for why they are the focus population and describe the risks associated with their involvment as participants

Does the research involve normal educational practices?

No

Is the research being conducted in an accepted educational setting?

No

Are participants in a class taught by the principal investigator?

No

Are the research activities part of the required course requirements?

No

Will course credit be offered to participants?

No

Amount of credit?

No

If class credit will be given, list the amount and alternative ways to earn the same amount of credit.Note: the time it takes to gain the same amount of credit by the alternate means should be commensurate with the study task(s)

How will permission to conduct research be obtained from the school or district? IRB approval is not permission to conduct the research. You need to access a gatekeeper. If you are implementing a survey with NC State populations, please make sure you follow the NC State survey regulation.

Will you utilize private academic records?

No

Explain the procedures and document permission for accessing these records.

Employees?

Yes

Describe where (in the workplace, out of the workplace) activities will be conducted.

The study will be conducted virtually with users who are employees of the NC Department of Transportation.

From whom and how will permission to conduct research on the employees be obtained?

From both the North Carolina Department of Transportation (the employer) and the individual employees.

How will potential participants be approached and informed about the research so as to reduce any perceived coercion to participate?

The NCDOT will provide a list of eligible employees. We will then solicit a random subset of this list to participate in the research. They will be informed this is optional and not required in any way and the sponsor will not know which employees we contact or who agreed to participate. Contact information for employees is public but not which employees are project managers or related staff.

Is the employer involved in the research activities in any way?

Yes

Please explain:

The employer (NCDOT) is the one sponsoring the research and has had input on the design of the survey. Participants are not directly affiliated with the funding unit within NCDOT that is associated with this study.

Will the employer receive any results from the research activities (i.e. reports, recommendations, etc.)?

Yes

Please explain. How will employee identities be protected in reports provided to employers?

No directly identifying information about the employees will be included in the report. Only aggregate statistics and carefully de-identified interaction patterns with the systems. If our N is too small and the aggregate statistics become re-identifiable, we will omit the re-identifiable portions from the report.

Impaired decision-making capacity/Legally incompetent?

No

How will competency be assessed and from whom will you obtain consent?

Mental/emotional/developmental/psychiatric challenges?

No

Identify the challenge and explain the unique risks for this population.

Describe any special provisions necessary for consent and other study activities (e.g., legal guardian for those unable to consent).

People with physical challenges?

No

Identify the challenge and explain the unique risks for this population.

Describe any special provisions necessary for working with this population (e.g., witnesses for the visually impaired).

Economically or educationally disadvantaged?

No

Racial, ethnic, religious and/or other minorities?

No

Non-English speakers?

No

Describe the procedures used to overcome any language barrier.

Will a translator be used?

No

Provide information about the translator (who they are, relation to the community, why you have selected them for use, confidentiality measures being utilized).

Explain the necessity for the use of the vulnerable populations listed.

Our system is built for the use of project managers and other project staff at the NCDOT. Thus this population is the only one suitable for the evaluation of our system because we believe that no other population has the necessary expertise (transportation construction project management) and sufficient familiarity with the NCDOT specifically.

State how, where, when, and by whom consent will be obtained from each participant group. Identify the type of consent (e.g., written, verbal, electronic, etc.). Label and submit all consent forms. Adult Non-Exempt Consent Form Template Exemption Consent Form Templates

Verbal consent will be obtained from the participant by a member of the research team before the participant begins the

experiment before being presented with any other experiment materials. We are requesting a waiver of signed consent because this research poses minimal risk to participants.

If any participants are minors, describe the process for obtaining parental consent and minor's assent (minor's agreement to participate).Parent/Guardian Permission FormMinor Assent Forms

Are you applying for a waiver of the requirement for consent (no consent information of any kind provided to participants) for any participant group(s) in your study?

No

For each participant group that you are requesting a waiver of consent for, please state what method this waiver is needed for, why it is needed and address each of the above 5 criteria to justify why your study qualifies for a waiver of consent.

Are you applying for an alteration (exclusion of one or more of the specific required elements) of consent for any participant group(s) in your study?

Yes

Identify which required elements of consent you are altering, describe the participant group(s) for which this waiver will apply, and justify why this waiver is needed.

I am altering the consent process by asking for broad consent after completing the interview with a participant. This

alteration of consent affects all research participants. I cannot do this research without an alteration of consent because we believe participants will have a better understanding of what broad consent will entail after their data has been collected. The research is no more than minimal risk because it involves typical tasks the participants which is already available to their employer and involves only

answering survey questions and interacting with a website. The alteration of consent won't affect the rights and welfare of participants because they will still be able to decline broad consent, just after data collection instead of before.

Are you applying for a waiver of signed consent (consent information is provided, but participant signatures are not collected)? A waiver of signed consent may be granted only if: The research involves no more than minimal riskThe research involves no procedures for which consent is normally required outside of the research context.

Yes

Would a signed consent document be the only document or record linking the participant to the research?

No

Is there any deception of the human subjects involved in this study?

No

Describe why deception is necessary and describe the debriefing procedures.Does the deception require a waiver or alteration of informed consent information?Describe debriefing and/or disclosure procedures and submit materials for review.Are participants given the option to destroy their data if they do not want to be a part the study after disclosure?

For each participant group please indicate how many individuals from that group will be involved in the research. Estimates or ranges of the numbers of participants are acceptable. Please be aware that participant numbers may affect study risk. If your participation totals differ by 10% from what was originally approved, notify the IRB.

3-10 NCDOT employees

How will potential participants be be found and selected for inclusion in the study?

The NCDOT will provide a list of eligible employees. We will select a random subset from this list and solicit them directly via email.

For each participant group, how will potential participants be approached about the research and invited to participate? Please upload necessary scripts, templates, talking points, flyers, blurbs, and announcements.

Participants will be contacted by the research team directly via email. The key talking points are:

We are testing a new version of the CLEAR website.

Researchers from NCSU request your participation in the study.

You are not required by the NCDOT to participate and if you refuse it will not be recorded or considered negatively on your employment. We will not inform them of your decision and they will not know you have been asked.

The study will take approximately 1 hour 30 minutes.

Please reach out directly to the NCSU research team if you would like to participate.

Describe any inclusion and exclusion criteria for your participants and describe why those criteria are necessary (If your study concentrates on a particular population, you do not need to repeat your description of that population here.) Inclusion and exclusion criteria should be reflected in all of your recruitment materials and consent forms.

Our inclusion criteria are adult (18+ years old) NCDOT employees who agreed to participate in the study and who agree

to have their voice and screen recorded. They need to be either construction project managers or related staff at the NCDOT. All others are excluded.

Is there any relationship between researcher and participants - such as teacher/student; employer/employee?

No

What is the justification for using this participant group instead of an unrelated participant group? Please outline the steps taken to migitate risks to participants from the pre-existing relationship, including power dynamics of this relationship and/or perceived coercion.

The research will be conducted by NCSU, but the NCDOT is the sponsor. As previously stated, we are targeting experts

within the NCDOT explicitly because they will be the end-users of the system and have the necessary expertise. The participants will be the only NCDOT representatives present for the experiment we will make it clear only anonymized interactions and statistical data will be presented to the NCDOT.

Describe any risks associated with conducting your research with a related participant group.

Describe how this relationship will be managed to reduce risk during the research.

How will risks to confidentiality be managed?

Address any concerns regarding data quality (e.g. non-candid responses) that could result from this relationship.

We are aware that participants may work harder than normal circumstances but have not other concerns since the NCDOT sponsors will not be present.

In the following questions describe in lay terms all study procedures that will be experienced by each group of participants in this study.For each group of participants in your study, provide a step-by-step description of what they will experience from beginning to end of the study activities. Should you prefer, you can upload a detailed study procedure packet and refer us to that document in this text box. If you choose to upload a procedures packet, do not discuss procedures in the below text box.

After preliminarily agreeing to participate, participants will receive a link to a video call and a scheduled date/time. After

joining the call they will be asked to keep their video camera off for the duration of the experiment.

Interview

We will collect consent verbally.

We will provide the written instructions to the participant (see attached).

Participants will be asked to begin "thinking-aloud" for the duration of the interview. They will be prompted to continue if they remain silent for more than 15 seconds. The prompt we will one of "Please continue to think out loud" or "Please continue to remember to talk about what you are doing and what you are thinking as you complete this task."

We will verbally give the pre-questionnaire.

First Condition

Participants will search the clear website to select lessons learned and best practices according to the instruction sheet. They will use the search functionality to find lessons learned and best practices related to their current project with the NCDOT. For every lessons learned or best practice they choose to open, they will be verbally given the per-ll-bp-questionnaire.

After 20 minutes or earlier if the participant chooses, this condition will end. They will be reminded of time remaining at approximately 15, 10, 5, and 1 minute remaining.

Second Condition

Participants will search the new AI-assisted website to select lessons learned and best practices. They will search in the same manner as steps 5 and 6. For every lessons learned or best practice they choose to open, they will be verbally given the per-ll-bp-questionnaire.

(Same as 6.) After 25 minutes or earlier if the participant chooses, this condition will end. They will be reminded of time remaining at approximately 15, 10, 5, and 1 minute remaining.

Wrap-up

We will verbally give the post-questionnaire.

Participants will be allowed to provide any additional feedback they may choose.

Participants will be asked if they wish to provide broad consent. This process will include thoroughly reviewing the broad consent addendum before requesting

Are you requesting the use of secondary information to be used as data for this research project? The secondary information can either currently exist or be generated in the future.

Discuss the following: permission to access the information (direct permission from the participant or records release), how researchers will access, transfer, store, and destroy the data.

Discuss the identifiable/re-identifiable nature of the data through either direct IDs, indirect IDs, or triangulation of datasets, data points, researcher access/expertise, or analysis. .

List all data categories to be requested (ex: age, race, student ID, GPA, ACT, Medical ID, diagnosis). Discuss if the data requires a Data Use Agreement

Discuss if the data are subject to FERPA, HIPAA, or the GDPR.

We are requesting the NCDOT provide a list of eligible staff for the study. These are construction project managers and related staff at the NCDOT. Only names are needed for the list since contact information is publicly available. This list is needed because it is not public who the project managers and related staff are and we will not be soliciting participation from the entire NCDOT. The only purpose will be selecting a random subset to contact for participation. The random subset contacted will not be shared with the NCDOT. It will be provided via a secure portal for NCDOT employees and contractors to which the research team has access. We will not store any local copies of this list and it will remain in the hands of the NCDOT to be destroyed or access revoked at their discretion.

Social/Reputational?

No

Academic (affect grades, graduation)?

No

Employment (affect job)?

No

Financial (affect financial welfare)?

No

Medical (harm to treatment)?

No

Insurability (harm to eligibility)?

No

Legal (reveals unlawful behavior)?

No

Private behavior (harm to relationships/reputation)?

No

Religious Issues/Beliefs?

No

Describe the nature and degree of risk that this study poses. Describe the steps taken to minimize these risks. You CANNOT leave this blank, say 'N/A', none' or 'no risks'. You can say "There is minimal risk associated with this research." For each 'Yes' selected above, describe the probability of the risk occuring and the magnitude of harm should the risk occur. Discuss how you are mitigating those risks through participant selection, study design, and data security.

There is minimal risk associated with this research because the participants routinely perform similar tasks for their

employer.

If you are accessing private records, describe how you are gaining access to these records, what information you need from the records, and how you will receive/record data. Private records may include: educational, medical, financial, employment. Some of these private records may be subject to laws such as FERPA and HIPAA. Your content here should match what you've discussed on the procedures tab.

N/A

Are you asking participants to disclose information about other individuals (e.g., friends, family, co-workers, etc.)?

No

You have indicated that you will ask participants to disclose information about other individuals (see Populations tab). Describe the data you will collect and discuss how you will protect confidentiality and the privacy of these third-party individuals.

If you are collecting information that participants might consider personal or sensitive or that if revealed might cause embarrassment, harm to reputation or could reasonably place the subjects at risk of criminal or civil liability, what measures will you take to protect participants from those risks?

N/A

If any of the study procedures could be considered risky in and of themselves (e.g. study procedures involving upsetting questions, stressful situations, physical risks, etc.) what measures will you take to protect participants from those risks?

N/A

Describe the anticipated direct benefits to be gained by each group of participants in this study (compensation is not a direct benefit).

There are no anticipated direct benefits to be gained by participants in this study.

If no direct benefit is expected for participants describe any indirect benefits that may be expected, such as to the scientific community or to society. The information presented in the study is directly relevant to their jobs and thus may improve their performance. However, this data is already freely available to them. If the project is successful, they would benefit by receiving a

better website for the NCDOT.

Will you be receiving already existing data without identifiers for this study?

No

Will you be receiving already existing data which includes identifiers for this study?

Yes

Describe how the benefits balance out the risks of this study.

Will data be collected in a way that would not allow you to link any identifying information to a participant?

No

Will any identifying information be recorded with the data (ex: name, phone number, IDs, e-mails, etc.)?

Yes

Will you use a master list, crosswalk, or other means of linking a participant's identity to the data?

Yes

Will it be possible to identify a participant indirectly from the data collected (i.e. indirect identification from demographic information)?

Yes

Audio recordings?

Yes

Video recordings?

Yes

Digital/electronic files?

Yes

Paper documents (including notes and journals)?

No

Physiological Responses?

No

Online survey?

Yes

Restricted Access (who, what, when, where)?

Yes

Password Protection (files, folders, drives, workstations)?

Yes

Suggestion of anonymous browsing?

No

Locks (office, desks, cabinets, briefcases)?

No

VPN (transfer, upload, download, access)?

Yes

Encryption (files, folders, drives)?

Yes

Describe all participant identifiers that will be collected from each data collection method (surveys, interviews, focus groups, existing data, background data collected via host site or software). Discuss why it is necessary to record identifiers at all and describe the deidentifying process

We will contact participants via email. We will not keep any information linking the contact to their data. After conducting all interviews, we will scrub timestamps from files. Consent will be collected verbally then we will record participants' screens and audio during the interview.

Participants will be asked to browse the NCDOT CLEAR webpage and our alternate website. The NCDOT site records IP addresses and pages visited, but each employee already has access to this site and should be using it somewhat regularly as part of their work. Our website does not store this information.

The screen and audio recordings will be used to code the entire interview. After an interview has been coded the screen and audio recordings will be discarded. No direct identifiers will be recorded for the questionnaires but they will be linked to the recordings. In addition, the questionnaire includes indirect identifiers as the duration of being a project manager and how many projects they have worked on. This information is necessary to evaluate the performance of our system across expertise levels but will only be reported with direct identifiers stripped.

If recording identifiable information about participants, discuss any links between the data and the participants and why you need to retain them. Discuss destruction of links or removal of identifers.

After coding the interview the corresponding audio and screen recordings will be destroyed. The surveys will be given

verbally during the interview so no additional identifiers will be recording there. Some of the questions could potentially be re-identifiable (namely years of experience and number of projects). I will also have a master list, which I need to track participation and code data. I will destroy the master list at the end of the research project.

Discuss if you'll be working with your departmental IT to create a data management plan and if you're using NC State managed devices, NC State Google Drive or other NC State non-networked device. If using a personal device, discuss data protection.

We will be storing data in NC State Google Drive shared only with the research team and not NCDOT. A mix of personal

and NCSU managed devices will be used to access the data. All team members understand that any device used to access the data must have all current security updates, have malware protection software, and comply with NCSU regulations regarding those updates, security software, and password requirements.

NC State Google Drive data is encrypted and only accessible by the team. It requires 2-factor authentication (via NCSU Shibboleth) and is password protected. No data will be stored outside of NC State Google Drive.

Data may also be stored on personal machines. These will be required to be password protected (using a strong password) with an appropriate security policy (e.g., lock screen timeouts). The individuals using personal machines will commit to meeting NCSU security policies for personal machines. The data must be kept in encrypted folders or files that require either a password, certificate, or preferably both to access.

Copies of the data may only be retrieved from the NC State Google Drive so that a common access log is maintained. If any copy is kept on a personal machine they must inform the PI.

Audio and screen recordings will be done on an NC State Zoom account and transferred using VPN to NC State Google Drive, where it will be kept for data analysis purposes and then destroyed in alignment with OIT best practices.

After each interview, we will delete our (email) correspondence with that individual. After coding the interview from the audio/screen recordings, those recordings will be deleted.

Describe any ways that participants themselves or third parties discussed by participants could be identified indirectly from the data collected, and describe measures taken to protect identities. (Data can be reidentified by researcher access, technology employed, researcher expertise, and triangulation of data or other information. Discuss the probability of reidentification and the magnitude of harm to participants should the data be reidentified. Discuss the probability of reidentification occuring and the magnitude of harm should it occur).

The survey question about prior experience could potentially be used to identify participants' responses. This is needed

because we believe more experienced participants will yield higher quality across conditions and that our system will most benefit those will less experience. Based on the responses we will take appropriate steps to bin the responses if they are unique enough to be identified. We do anticipate minimal harm if a participant's responses were identified, since the task involved is very similar to their normal work routine which is public to their employer.

For all recordings of any type:Describe the type of recording(s) to be made Describe the safe storage of recordings Who will have access to the recordings? Will recordings be used in publications or data reporting? Will images be altered to de-identify?Will recordings be transcribed and by whom?

We will conduct the experiment via a zoom call and we will record the screen for interactions with the websites and audio for a think-aloud protocol. Participants will be asked to keep their cameras off to avoid video recording but are welcome to keep the video on if they are comfortable to do so. The video will be helpful for us as communication can be non-verbal and we can capture data that we wouldn't otherwise be able to. All recordings (audio, screen recordings, and in some cases video) will be discarded after the responses are coded into interaction patterns with direct IDs stripped. Only the coded interactions will be used in any publications or reports. All coding will be done by the NCSU research team and the recordings will never be shared with the NCDOT.

Describe how data will be reported (aggregate, individual responses, use of direct quotes) and describe how identities will be protected in study reports. Reporting data may sometimes reidentify your participants. If needed, you can adjust how you report your data to protect the identities of your participants. Discuss.

Data will be reported through aggregate statistics (demographic Ns will only be reported in N of 3 or greater) and coded

interactions with direct identifiers stripped.

Will anyone besides the PI or the research team have access to the data (including completed surveys) from the moment they are collected until they are destroyed? This includes sharing data with sponsors, journals, or using the data for future research endeavors. If you are sharing the data, this should be in your consent form.

Only the research team will have access. We may use the de-identified data (which is re-identifiable due to small overall

N and the nature of the data) aggregate statistics and interaction patterns in further research if we have broad consent from participants. Select data will be shared with the NCDOT and published, which is disclosed in the consent form. We will not keep, reuse, or share the audio/screen/video recordings.

Describe any compensation that participants will be eligible to receive, including what the compensation is, any eligibility requirements for that compensation, and how that compensation will be delivered. Examples of compensation include: monetary compensation, research credits, raffle/drawing, novel items. Make sure to check with your department regarding issues of tracking payments as your department accounting office may have requirements that affect your human subjects privacy (such as the mandatory tracking of anyone who receives compensation). This tracking may influence the confidentiality/anonymity of your research and must be addressed in this application.

No compensation will be provided.

Explain compensation provisions if the participant withdraws prior to completion of the study.

No compensation will be provided and there are no penalties for withdrawal prior to completion.

Appendix B. Study Instructions

Instructions for Participants

Thank you for choosing to participate in this study. The study will begin with a questionnaire to assess your experience as a project manager and your particular familiarity with your current project. Then we will begin the interactive portion of this study.

1. After completing the questionnaire, you will browse the CLEAR website to search for lessons learned and best practices related to your current project. You will have 20 minutes to complete this task. You may finish the task early if you choose.

2. Next, you will be presented with a new list of keywords to repeat the task. However, this time you will use a new version of the website. This will once again be a 20-minute session, but you may choose to stop sooner.

This will be a "think-aloud" study. That means we request you to continually tell us what you are thinking while completing tasks. In addition to thinking-aloud, we will prompt you to evaluate the lessons learned and best practices you choose to view. If you stop thinking-aloud we will prompt you to continue.

The last portion of this study will be a post-questionnaire to ask about your general preferences for the 2 tasks you performed, and we will give you time to provide any verbal feedback you wish to be included in the study.

Appendix C. Questionnaires for the Study

Pre-Survey Questionnaire

1. What is your current role?

2. How long have you been a project manager or related staff?

3. Approximately how many projects have you worked on as a project manager or related staff?

4. Please rate your expertise as a project manager or related staff.

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Somewhat confident) (Confident) (Very Confident)

5. Approximately how many projects similar to your current project have you worked on as a project manager or related staff?

6. How would you rate your expertise on this type of project?

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Somewhat confident) (Confident) (Very Confident)

7. Please indicate how often you use the CLEAR website.

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Rarely) (Somewhat frequently) (Very Frequently)

Post-Survey Questionnaire

1. Please rate the relevance of the current CLEAR website search results.

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Low) (Moderate) (High)

2. Please rate the relevance of the new website search results.

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Low) (Moderate) (High)

3. Please rate the overall quality of the LL/BPs in the CLEAR database.

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Low) (Moderate) (High)

4. Please indicate which of the two tasks you preferred.

   - Current CLEAR website search
   - New website search

Per-LL/BP Questionnaire

1. Have you read this LL/BP before?

2. Please rate the relevance of this LL/BP to your current project.

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Not Relevant) (Relevant) (Very Relevant)

3. Please rate your familiarity with the subject of this LL/BP.

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Not Familiar) (Average) (Very Familiar)

4. Please rate the quality of this LL/BP.

1 ---------- **2** ---------- **3** ---------- **4** ---------- **5**

(Low) (Moderate) (High)