Lauble, S., Zielke, P., Chen, H. & Haghsheno, S. (2023). Process analysis with an automatic mapping of performance factors using natural language processing (NLP). *Proceedings of the 31st Annual Conference of the International Group for Lean Construction (IGLC31),* 59–68. doi.org/10.24928/2023/0140

PROCESS ANALYSIS WITH AN AUTOMATIC MAPPING OF PERFORMANCE FACTORS USING NATURAL LANGUAGE PROCESSING

Svenja Lauble[1], Philipp Zielke[2], Hongrui Chen[3], and Shervin Haghsheno[4]

ABSTRACT

In lean construction projects, much information is collected during the process analysis with the trades. This data is increasingly documented as a reference for use in future construction projects. By doing this, efficient methods are required to use this data. Often, the unstructured naming of data is a challenge for a rule-based allocation of information, and manual work is required to identify the needed data. Therefore, the aim is to develop an automatic mapping of historical performance factors to the tender specifications of a new construction project. To support the process analysis with historical project data, a case study is executed using Natural Language Processing (NLP). With a NLP model, the process descriptions from the tender specifications of the new construction project can be compared with a master database, to filter the right performance factor and calculate the duration for a process. This procedure can be used to support the further process analysis together with the trades to generate a validated construction schedule. The case study shows promising results in the prediction results. First, the mapping quality and second, the prediction accuracy are evaluated. With the developed mapping concept, last planners can validate their estimations of durations in lean construction process planning with a target to support stability in a project. Still, a more detailed description of the processes could increase the prediction results.

KEYWORDS

Digitization, lookahead planning, work structuring, process, complexity.

INTRODUCTION

Regarding the continuous improvement process (CIP) and a well-founded knowledge management system in lean construction projects, companies tend to collect more and more data about their projects. According to a survey by Thomas and Bowman (2022) with 3.916 stakeholder in the construction industry, their data has doubled within three years (from 2019 to 2022). In lean construction projects, much information is documented about the processes. In the Last Planer System and in Takt planning (Haghsheno 2016) a process analysis is done together with the trades as knowledge carriers. A sequence of processes for each product type is compiled here, along with the time and resources required for each process step. This information is increasingly documented as a reference and updated with the real data during project execution. For example, Choo et al. (1998) collect weekly updated data of construction processes using a database called "WorkPlan". This data is then used to update the weekly work planning of the respective project. By documenting the duration of a work package and the

[1] Research Assistant, Karlsruhe Institute of Technology, Germany, +49-721-608-41513, svenja.lauble@kit.edu
[2] Research Assistant, Karlsruhe Institute of Technology, Germany, +49-721-608-41518, philipp.zielke@kit.edu
[3] M. Sc. Student, Karlsruhe Institute of Technology, Germany, +49-721-608-43650, hongrui.chen@student.kit.edu
[4] Professor, Karlsruhe Institute of Technology, Germany, +49-721-608-42646, shervin.haghsheno@kit.edu

manpower behind it, realistic performance factors can be documented (Haghsheno 2016). Figure 1 displays an exemplary database for lean construction projects. In this database, processes are described. With their durations and the manpower available, realistic performance factors can be calculated. A performance factor defines the required calculation time for one unit of a process task with one person. These performance factors, together with the process description, can serve as a master database for the process analysis of a new project (see Haghsheno et al. 2016).

For example, Frandson et al. (2013) define 'gathering information' about the process as the first phase of the Lean method's Takt planning. And, for the Last Planner System, the phase and lookahead planning with the trades can be supported by historical information. Here, the master database can serve as a starting point for planning and as basis for discussions with the construction trades.

Database for Lean Construction projects

Process name	**Duration (D)** (in hours)	**Manpower (M)**	**Units (U)**	**Performance factor (D/M/U)** (hours / unit)	**Comment (e.g. special conditions)**
XXX	8	4	10	0.2	Snow
XXX	6	2	40	0.075	-
…	…	…	…	…	…

Figure 1: Database example for Lean Construction projects.

As a result, unnecessary buffers and capital commitment costs for construction employees and their machines are reduced. Also, time pressure can be prevented by enhancing motivation, security, and quality (Rogel 2013).

With detailed work steps, covering different types of construction, and the existing product complexity in construction, databases can get very complex and large. This is also shown by publicly available databases. E.g., the Construction Cost Information Center for Architects in Germany, called 'BKI' (Baukosteninformationszentrum), is documenting several thousand possible work steps with their average performance factor and costs per process description. Besides complex databases, a mapping to process descriptions of new construction projects can get quickly complex, as often the naming is not standardized. Following a manual mapping of the correct performance factors causes high levels of manual work with the danger of a misallocation. With the prospect of manual effort, a detailed use of master databases in a process analysis can be prevented.

This paper therefore has the target of analyzing how an automatic mapping of performance factors from a master database to the process description of a new construction project can be designed. Figure 2 displays the concept of the automatic mapping with the master database (right) and the process descriptions of a new lean construction project (left). For example, the tender specifications can serve as the basis for the process descriptions. By mapping the processes of the new project with those of the database, a suitable performance factor is identified. Together with the units, a duration for each process can be calculated.

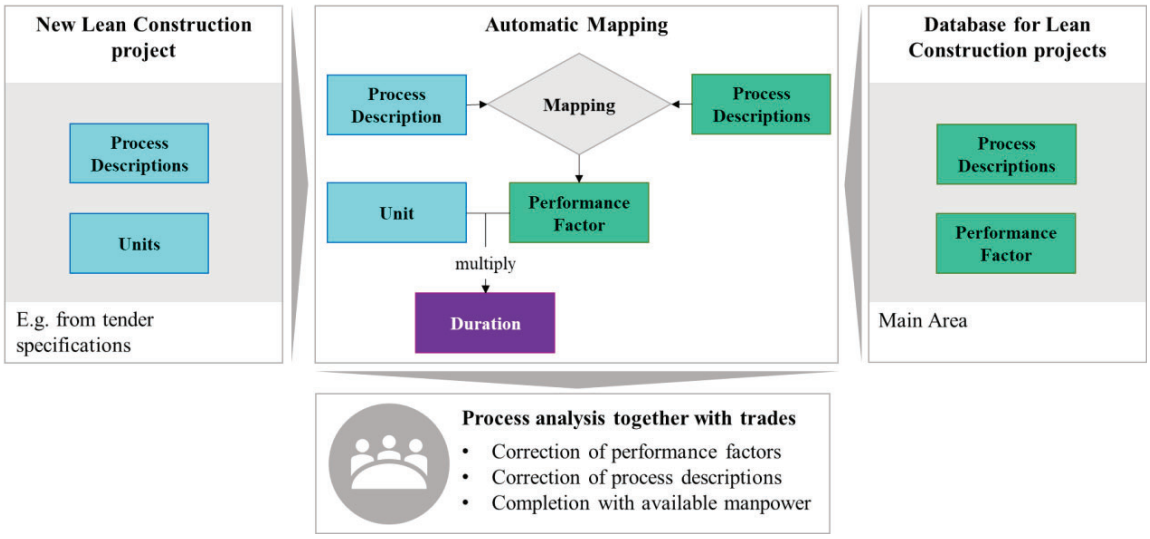

Figure 2: Concept of an automatic mapping of performance factors for a process analysis.

For the automatic mapping, the method of 'Natural Language Processing' (NLP) is first described and then evaluated in a case study.

THEORETICAL FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

According to Cambria and White (2014), NLP is a theory-motivated range of computational techniques for the automatic analysis and representation of human language. IBM (Education, 2020) considers NLP to be a branch of Artificial Intelligence (AI). It is concerned with the ability to enable computers to understand text and spoken words in a way that is almost identical to that of humans. Liddy (2001) provided a detailed definition: "NLP is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications". EasyAI (2019) explained NLP concisely and in an easy-to-understand way: "NLP is the bridge that communicates between machine language and human language." Language is typically unstructured data. NLP is used to let the machine understand and use this information. A consensus can be seen: with the support of NLP, humans can be supported in the lean construction process planning.

So far, NLP has already been part of pilot studies in construction research. Jagannathan et al. (2022) apply NLP to analyze unstructured text data in annual reports of construction contracting companies. Li et al. (2020) use NLP to predict the probability of obtaining construction accident compensation through a practical example in Hong Kong. Wang et al. (2022) develop a virtual assistant with the use of NLP, with whom information retrieval for construction project team members is supported.

According to EasyAI, Deep Learning-based NLP can be divided into three steps: corpus pre-processing, design modeling, and model training. For the pre-processing, the following six tasks can be executed (EasyAI 2019, Bachani 2020):

Tokenization: Tokenization is the breakdown of long texts like sentences, paragraphs, and articles into word-based data structures for subsequent processing and analysis work.

Stemming: Stem extraction is the process of removing the prefixes and suffixes of words to get the root word. Common prefixes and suffixes are "plural of noun", "progressive", and "past participle". For example, "playing" will be converted to "play".

Lemmatization: Lexical reduction is based on the dictionary and transforms the complex form of a word into its most basic form. For example, "drove" will be converted to "drive". Each language requires semantic analysis and parts of speech to establish a complete lexicon.

Parts of Speech: In traditional grammar, a part of speech is a category of words that have similar grammatical properties, such as "nouns", "conjunctions", and "verbs".

Named Entity Recognition (NER): NER, also known as "proper name recognition", refers to the recognition of entities with specific meanings in the text, mainly including names of people, places, institutions, proper names, etc.

Chunking: Chunking is the process of grouping the words in unstructured text and making up phrases.

Easily spoken to design the model, the pre-processed words are used in a **word embedding**. The purpose of word embedding is the vectorization of words. Each word or phrase is mapped to corresponding vectors of real numbers. With these dependencies, words can be allocated. Algorithms such as deep learning can ingest and process these vectors to formulate an understanding of natural language. (Collis, 2017)

Following, the process description of the master database as well as the process description of the new construction project, the data is pre-processed and afterwards vectorized. Hence, the tasks can be compared by their real numbers, and the effort value with the closest match is used as the planning basis for the process analysis. This procedure also serves as the foundation for the methodical approach.

CASE STUDY

METHODICAL APPROACH AND DATA

The methodical approach for the case study, is described in Figure 3. For the case study two datasets are used: the master database and the process description of tender documents from a real-world construction project. First, the mapping quality is evaluated with two indicators, and afterwards, the prediction accuracy for the duration is analyzed.

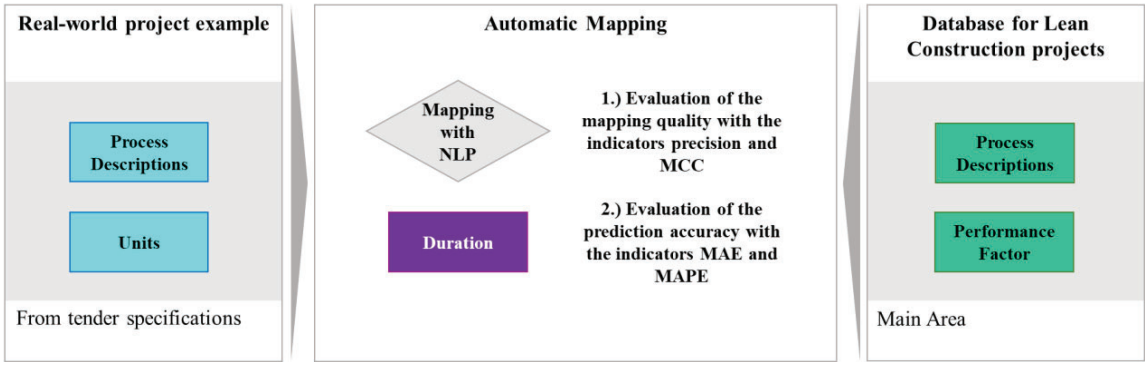

Figure 3: Methodical approach for the case study.

As a master database, the data from BKI is used for the research model. The available dataset contains 3.586 described process steps for new construction projects. The dataset is complete and free of missing data, and its reliability is high. The dataset is available in German and is annually updated. The dataset contains, after removing irrelevant features, six columns with: category type, short task description, long task description, unit, average cost per unit, and performance factor. The mean value of the short task description, text length are 34.29 digits,

for the long task description it is 383.06 digits. The mean value is defined as the sum of all digits divided by the number of projects. The performance factor presents the working time in hours, based on one unit. The average performance factor is 1.23 hours with a standard deviation of 9.16 hours, a minimum of 0.01 hours, and a maximum of 260 hours per unit.

The real-world construction project data contains process steps for structural engineering work (earthwork, concrete, and masonry work) described in the tender specifications. In the real-world project data, there is no information about performance factors included. This information is needed to start a well-founded lean construction process plan. The dataset includes 194 entries with a short process description as well as the number of units. To these two columns, the corresponding process description and the performance factor of the master database are manually matched. Entries with no match and null values are deleted. The performance factor is then multiplied by the number of units to calculate the duration. This results in 93 data entries with an average duration of 0.52 hours per unit (performance factor) and 87.60 hours per process step (duration per unit multiplied by the respective unit). Table 1 summarizes the mean and standard deviation for the master database and the real-world project for the performance factors and duration. The performance factors of the real-world project are on average lower than those of the master database and show a lower standard deviation.

Table 1: Description of the data

	Master database	**Real-world project**	
	Performance Factor	**Performance Factor**	**Duration**
Mean	1.23 hours	0.52 hours	87.60 hours
Standard deviation	9.16 hours	0.57 hours	290.06 hours

In the first phase, the master database is prepared with data cleaning process, encoding, and splitting. The data cleaning contains data truncation, data enhancement, and selecting a classification objective:

Data Truncation: Nevertheless, the available dataset of real project consists only of structural engineering work. To control the required resources within the limits of Google Colab and maximize model performance with restricted resources, the master database is limited to the category of structural engineering. After restricting the range, the number of data sets dropped from 3.586 to 1.420.

Data Enhancement: Despite its high quality, the BKI dataset presents a problem for machine learning. In general, to train a classifier, each class always needs multiple data samples for algorithm learning. However, there is only one unique sample to be classified in each class in the current BKI dataset. Therefore, the expansion of the data cannot be avoided, and a replication is performed. In a study, IBM researchers explored different classifiers, and the performances of the classifiers with different numbers of training samples were compared. The results show that the model improves significantly with the inclusion of ten samples (Anaby-Tavor et al., 2019). Therefore, the number of replications of the master database is set to 10. After the data expansion procedure, the number of data points in the dataset increased from 1,420 to 14,200.

Selecting Classification Objective: After filtering the process steps with effort values, there are 14,030 data points left.

Afterwards, the data is encoded. After investigating the duplicate entries, label encoding is used for the short description. Label encoding assigns a number starting from zero to each possible category in the short description column (Yadav 2019). The new dataset includes 14.030 data points and 1.366 category encodings.

In the last step of data preparation, the master dataset will be divided into a training set, a validation set, and a test set. This procedure is known as data splitting. Brownlee (2020) defines data splitting as a technique that is used to evaluate the performance of machine learning algorithms and can be employed in any supervised learning algorithm. The training and test sets were split by 80% and 20%.

After data preparation, the model is developed and various publicly available libraries for Python are utilized: TensorFlow, Transformers, Tune and Scikit-learn. For the model, the framework GBERT is used, which Chan et al. (2020) state the best performing German framework for NLP. The model further on uses several pre-defined hyper-parameters, such as a dropout rate of 0.1, 10 epochs, a batch size of 16, and a learning rate of 2e-5. For the model, the short and the long process descriptions are combined by the separator token "[SEP]".

After model training, the model will be evaluated by several metrics, such as:

Precision: Precision indicates the proportion of units that the model predicts to be positive and are in fact positive. The mathematical definition is the proportion of the true positives (TP, actual positive and labelled as positive) to the true and false positives (FP, actual negative and labelled as positive).

$$Precision = \frac{TP}{TP + FP}$$

MCC: Matthews Correlation Coefficient (MCC) represents the correlation between the true value and the predicted value. It ranges from -1 to 1. A score of 1 indicates a very good prediction, while a value close to 0 means that the model performs poorly and is like the random classification. The value -1 represents inverse prediction (Grandini et al., 2020). The metric focuses only on whether each class is well predicted, regardless of class imbalance (T, 2021). The formula of MCC with TP, FP, true negatives (TN, actual negative and labelled as negative) and false negatives (FN, actual positive, labelled as negative) is:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

MAE: To calculate the Mean Absolute Error (MAE) for each known output y and the associated predicted value ŷ, the loss is calculated from Loss (y, ŷ) = |y - ŷ | (Russell and Norvig 2012, Hyndman and Koehler 2006). This metric is characterized by a stronger robustness, and outliers do not affect the result as much. The MAE is calculated with (Sammut and Webb 2010):

$$\boldsymbol{MAE} = \frac{\sum_{i=1}^{n} |\boldsymbol{y_i} - \widehat{\boldsymbol{y_i}}|}{\boldsymbol{n}}$$

MAPE: The Mean Absolute Percentage Error (MAPE) serves as an indicator of how good the predicted duration is. The smaller the indicator, the higher the prediction accuracy. The MAPE is the ratio of the difference between the actual output value y and the predicted value ŷ to the actual output value y over all data points (Hyndmann and Koehler 2006).

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Results of the Case Study

To evaluate the performance of the mapping model, itis evaluated with the test dataset from the master database and the real-world dataset. The test dataset of the master database contains 2.806 data entries. The accuracy reaches 0.92. The MCC represents the correlation between the true value and the predicted value. With this metric, it is possible to observe whether each class is well predicted or not. Here, MCC is at 0.92. Overall, the performance of the model on the test set is reasonably positive.

For the real-world dataset, the accuracy reaches 0.65. The value of MCC is 0.64. It is possible to conclude that the model performs less well on the dataset of a real project than on the test set. For both, the metrics are summarized in Table 2.

Still, the matching effort values can be very close, as the assigned tasks can be similar and the matched performance factor is close to the correct one.

Table 2: Overview of the mapping quality

Indicator	Value	
	Master database	Real-World dataset
Accuracy	0.92	0.65
MCC	0.92	0.64

Therefor the manually mapped durations and the predicted ones are compared for the real-world dataset. As indicators the MAE and MAPE are used as metrics. The MAE is 48.75 hours and the MAPE is 17.63 % (see Table 3).

Table 3: Overview of the prediction results

Indicator	Value
MAE	48.75
MAPE	17.63 %

While further analysing these metrics with a histogram shown in Figure 4, there are five mismatches with a difference of more than 50 hours between the predicted and actual duration. The highest mismatch is 2.340 hours, a high outlier. The Q3-quantil shows that 75 % have a difference of 2.5 hours between the predicted and actual value. In 85 % of the mismatches the difference is less than 8 hours or a working day (with 8 hours). The distribution of the prediction errors is displayed in Figure 4.

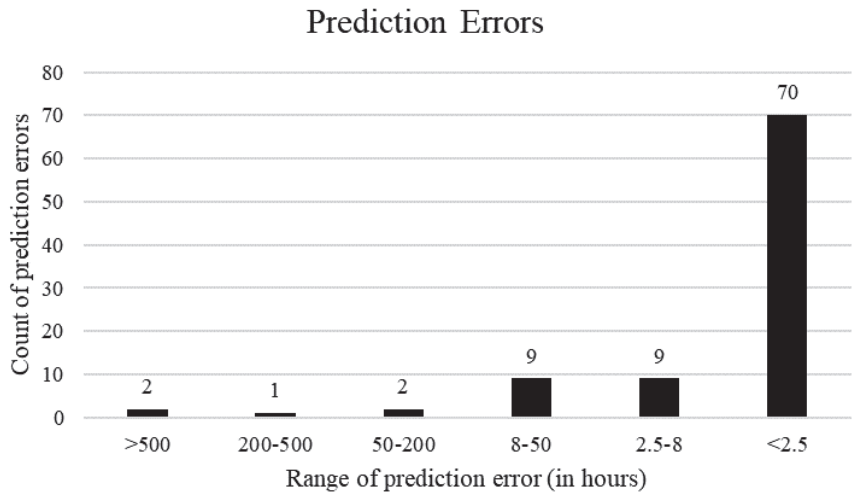

Figure 4: Distribution of Prediction Errors by hours.

Comparing the 2.5 hours in 75 percent of the cases to the 87.60 hours mean duration in the real-world dataset, the overall quality of the prediction is very good. Following, the predicted duration can serve as a basis for further planning. Due to the high deviations with high outliers, an expert must still be included in the scheduling with the target to reduce these outliers and correct the performance factors. These experts can be last planners, as they can validate the predicted durations best with their knowledge and, at the same time, challenge their own estimations.

CONCLUSIONS

In this paper, a concept was evaluated using a master database with historical project data to map the process descriptions to thoseof a new construction project. By doing this, the performance factor of the master database can be multiplied with the number of units of the construction project and a duration as the basis for planning is calculated. As the process description is often not standardized, the method of NLP is used. NLP serves as a bridge between machine language and unstructured human language.

For the concept evaluation, two databases were used. One master database with 3.586 entries describing new construction projects and the tender specifications of a real-world construction project with 194 entries. The mapping quality of the process descriptions between the master data base and the tender specification was 0.65 accuracy and 0.64 MCC. Both indicators show an average to good mapping quality. The higher mapping quality of the test dataset shows the need for a more detailed process description. When further comparing the prediction quality for the duration, the MAPE is 17.63 % and in 75 % of the mismatched cases, the prediction error is smaller than 2.5 hours. This deviation from the 87.60 hours of mean duration is very low, and the concept of mapping performance factors automatically shows its potential. With a first automatic mapping and prediction of the process durations, the time for scheduling can be accelerated and the quality of planning enhanced.

Figure 5 shows the full concept model with recommendations for the construction companies. The master database can be used to map with NLP models process description to the tender specifications of new construction projects. This information can serve as the basis for planning with the trades and their last planners. During the process analysis, performance factors and process descriptions are corrected, and the estimated durations are finalized with the available manpower. The result is a construction schedule that is validated by historical

project data. When documenting the performance of the project during its realization, the performance factors can be updated and detailed, and the master database can be enriched by new entries.

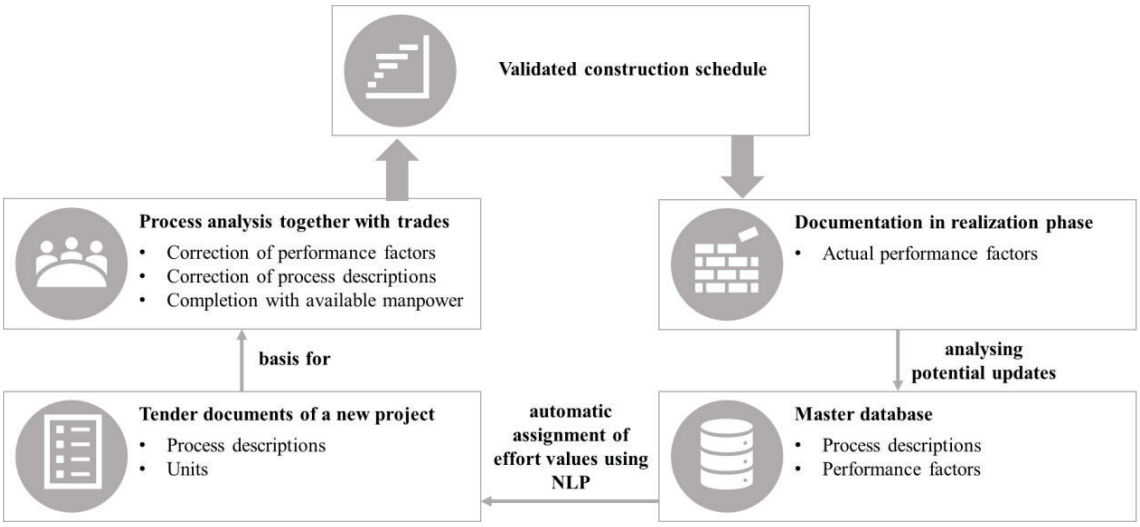

Figure 5: Concept model using NLP to automatically map performance factors to the tender specifications to support the process analysis.

As the case study is performed with one real-world dataset, there needs to be a broader field study using the tender documents of several construction projects. Also, the model acceptance of the schedulers and trades must be evaluated. Here, especially, the allocation of mismatches by the defined model should be analysed with the target of reducing the outliers and strengthening the collaboration between experts and machines.

REFERENCES

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., & Zwerdling, N. (2019). Not Enough Data? Deep Learning to the Rescue!, 5-6 http://arxiv.org/abs/1911.03118

Bachani, N. (2020). Chunking in NLP: Decoded. When I started learning text processing. . . | by Nikita Bachani | *Towards Data Science*. Retrieved September 1, 2022, from https://towardsdatascience.com/chunking-in-nlp-decoded-b4a71b2b4e24

BKI (2022). Baukosten: Planung und Daten für Architekten | BKI [Construction Cost Information Center for Architects in Germany]. https://bki.de/kostenplanung.html

Brownlee, J. (2020). Train-Test Split for Evaluating Machine Learning Algorithms. Retrieved September 1, 2022, from https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, 9 (2), 48. doi.org/10.1109/MCI.2014.2307227

Chan, B., Schweter, S., & Möller, T. (2020). German ' s Next Language Model. http://arxiv.org/abs/2010.10906

Choo, H. J., Tommelein, I. D., Ballard, G., & Zabelle, T. R. (1998). Workplan database for work package production scheduling. *In Proceedings sixth Annual Conference of the International Group for Lean Construction (IGLC 6)*.

Collis, J. (2017). Glossary of Deep Learning: Word Embedding. Retrieved September 1, 2022, from https://medium.com/deeper-learning/glossary-of-deep-learning-word-embedding-f90c3cec34ca

EasyAI. (2019). Understand natural language processing NLP in one article (4 applications+ 5 difficulties + 6 implementation steps). Retrieved September 1, 2022, from https://easyai.tech/en/aidefinition/nlp/

Education, I. C. (2020). What is Machine Learning? Retrieved September 1, 2022, from https://www.ibm.com/cloud/learn/machine-learning

Frandson, A., Berghede, K. & Tommelein, I. D. (2013). Takt Time Planning for Construction of Exterior Cladding In: Formoso, C. T. & Tzortzopoulos, P., *In Proceedings 21th Annual Conference of the International Group for Lean Construction (IGLC 21)*.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: An Overview, 10. doi: 10.48550/arXiv.2008.05756

Haghsheno, S., Binninger, M., Dlouhy, J., & Sterlike, S. (2016). History and theoretical foundations of takt planning and takt control. *In Proceedings of the 24th Annual Conference of the International Group for Lean Construction (IGLC 24)*.

Hyndman, R.J. & Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688.

Li, R. Y. M., Li, H. C. Y., Tang, B., & Au, W. (2020). Fast AI classification for analyzing construction accidents claims. *Proceedings of the 2020 Artificial Intelligence and Complex Systems Conference*, 1 – 4. doi: 0.1145/3407703.3407705

Liddy, E. D. (2001). Natural Language Processing, 15. Retrieved September 1, 2022, from https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub

Rogel, D., & Osebold, R. (2013). Skizzierung fachspezifischer Unsicherheiten im Bauwesen als Entwicklungsansatz zur Steigerung der Zeiteffizienz in der baubetrieblichen Terminplanung. Exploring Uncertainty: Ungewissheit und Unsicherheit im interdisziplinären Diskurs [Outlining discipline-specific uncertainties in construction as a developmental approach to increasing time efficiency in construction scheduling. Exploring Uncertainty: Uncertainty and Uncertainty in Interdisciplinary Discourse.], 209-244.

Russell, S. J., Norvig, P., & Davis, E. (2010). Artificial intelligence: A modern approach (3rd ed). *Prentice Hall*. https://zoo.cs.yale.edu/classes/cs470/materials/aima2010.pdf

Sammut, C. & Webb, G. I. (2010). Mean Absolute Error. In Encyclopedia of Machine Learning, *Springer US*, 652 – 652.

T, B. (2021). Comprehensive Guide on Multiclass Classification Metrics. Retrieved September 1, 2022, from https://towardsdatascience.com/comprehensive-guide-on-multiclass-classificationmetrics-af94cfb83fbd

Thomas, E., & Bowman, J. (2022). Harnessing the data advantage in construction. *AutoDesk & FMI report*.

Wang, N., Issa, R. R. A., & Anumba, C. J. (2022). Transfer learning-based query classification for intelligent building information spoken dialogue. *Automation in Construction*, 141, 104403. doi.org/10.1016/j.autcon.2022.104403

Yadav, D. (2019). Categorical encoding using Label-Encoding and One-Hot-Encoder. Retrieved September 1, 2022, from https://towardsdatascience.com/categorical-encoding-using-label-encodingand-one-hot-encoder-911ef77fb5bd