2024 EC³
2024 European Conference on Computing in Construction
Chania, Crete, Greece
July 14-17, 2024

METADATA EXTRACTION OF RFIs USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING ALGORITHMS

Ceyhun Ozogul[1], Esin Ergen[1]
[1]Istanbul Technical University, Istanbul, TURKEY

Abstract

Accurately assigning metadata of RFIs plays an important role in the analysis and management of RFI documents. However, these metadata are manually entered in the RFI management system, which results in loss of time and incorrect entries. This study aims to demonstrate that metadata of RFI documents can be extracted and assigned automatically using natural language processing and machine learning algorithms. To achieve this aim, the performance of Naïve Bayes and K-Nearest Neighbor algorithms are evaluated and compared. The results show that machine learning models perform well in automatically extracting the metadata of RFIs and, the performance of machine learning models for each label varies. The findings of this study can be used to develop an artificial intelligence based RFI management system by integrating natural language processing and machine learning models into the system.

Introduction

In construction projects, documents such as drawings and specifications may not address all aspects of the structures to be built or may contain deficiencies, uncertainties, and overlaps. These issues need to be clarified through Request For Information (RFI) documents (Shim et al., 2016). Hanna et al., (2012) defines RFI as " *a formal written procedure initiated by the contractor seeking additional information or clarification for issues related to design, construction , and other documents*". RFIs are a communication tool between the design team and the construction team (Bhat, 2017). Even though it holds significance, RFI, as a communication channel, is frequently perceived negatively within the project (Aibinu et al., 2019). This is due to the effort involved in initiating an RFI, as well as the process of reviewing and generating responses (Afzal et al., 2023). The response time for RFI can become critical and have a negative impact on progress in the field (Kelly and Llozor, 2020). Moreover, a delay in responding to the RFI, or the absence of a response altogether, can lead to a sense of frustration and mistrust among project team members (Philips-Ryder et al., 2012; Afzal et al., 2023).

Utilizing common data environments, such as Aconex and Procure, assists in overcoming the challenges associated with Request for Information (RFI) management. These platforms facilitate online communication and provide a means to track information effectively (Das, Tao and Cheng 2020; Afzal et al., 2023). However, metadata, which is a classification output, is typically entered manually into common data environments and the users are reluctant to fill out the related metadata, since this process requires additional time and effort. This results in inaccurate or incomplete metadata, which is problematic because metadata holds significant importance when analysing the unstructured RFI text data to obtain valuable insights and lessons learned to manage future projects in a more effective way. The erroneous metadata entry directly impacts the data quality, consequently influencing the outcomes of analysis studies.

This paper proposes to use Natural Language Processing (NLP) and machine learning models to automatically extract discipline code, which is one of the RFI metadata. To achieve this, NLP processes were applied to the RFI documents and Naïve Bayes (NB), K-Nearest Neighbor (KNN) algorithms were used to classify the RFI documents according to their disciplines. This approach enables extraction of the metadata of RFIs in an automated way. It reduces the time required for filling out the metadata of RFIs and results in effective management and analysis of RFIs by reducing erroneous or incomplete metadata.

Literature Review

Automatic extraction of metadata from documents is associated with the automatic classification of documents. Documents can be classified according to their metadata by using NLP and machine learning models. In the early 2000s, studies were conducted on the automatic extraction of metadata from texts in different domains, such as marketing, information-document management, educational sciences (Paik et al., 2001; Han et al., 2003; Yılmazel and Finneran, 2004). In these studies, rule-based methods were combined with NLP. Later, with the increase in the processing power of computers, machine learning models have been integrated with NLP instead of rule-based approach to automatically extract metadata.

Valdez et al. (2016) developed a natural language processing-based ontology for extracting metadata from texts in the biomedical domain. In addition, a natural language processing-based ProvCaRe-NLP tool was developed to perform clinical text analysis and information extraction. Some other studies used NLP and ML on metadata stored in a computerized maintenance management system (CMMS). Zhang et al. (2020) employed the information contained in failure notifications to forecast failure codes and assess the

precision of the labels. Texts are converted into word counts after NLP preprocessing. The SMOTE (Synthetic Minority Oversampling Technique) algorithm is utilized to balance the classes. In the final stage, one of the eight supervised machine learning models that are used in the classification problem, such as Support Vector Machine (SVM), NB, or Logistic Regression (LR), was applied.

Arif-Uz-Zaman et al. (2017) sought to enrich the metadata in a Computerized Maintenance Management System (CMMS) by classifying downtime notifications as either resulting from a failure or indicating no failure. In this study, the vectors obtained after applying NLP operations to the text data are used as input for SVM and NB algorithms. The SVM algorithm is noted to achieve the highest performance.

Tanguy et al. (2020) applied the SVM algorithm to categorize fault notifications in their study. The study demonstrated that incorporating all 3-character substrings along with word stems resulted in a slight enhancement compared to using only words. This improvement is likely attributed to the inclusion of abbreviations, compound terms, and alternative writing forms.

Deloose et al. (2023) have made a recent contribution to the study of CMMS metadata, in which natural language processing techniques and artificial models are utilized to forecast and rectify CMMS metadata. The performance of shallow machine learning models and deep machine learning models using multiple labels were compared. Random forest algorithm (RF), which is a shallow machine learning model, and Recurrent Neural Networks (RNN), which are deep learning methods, gave better results than other algorithms. It was also identified that the RNN algorithm performed slightly better than the RF algorithm.

In the field of construction management, studies on automatic classification of documents focuses on the accident analysis/occupational safety management, contract management, building information modeling (BIM), and facility management. In the studies related to the accident analysis and occupational safety management, accident reports were analyzed and the performances of machine learning models in the automatic classification of accident reports were compared. Tixier et al. (2016) used Stochastic Gradient Tree Boosting (SGTB) and Random Forest (RF) algorithms. The SGTB algorithm performed better than the RF algorithm. In Goh and Ubeynarayana (2017), the performances of six different machine learning models were compared, and the SVM algorithm showed the best performance.

In the area of contract management, studies involving automatic classification were conducted using NLP and ML (Yilmaz and Dikbas, 2013; Yilmaz, 2013; Candas, 2022; Eken, 2022). In Yilmaz and Dikbas (2013), dispute decision documents were classified using four machine learning algorithms. The algorithm with the highest accuracy was decision tree (DT). Eken (2022) conducted classification work to automatically review construction contracts. Five different machine learning algorithms are combined with different vectorization techniques. Models that perform well were selected and combined with ensemble learning. Candas (2022) conducted research on the multi-obejctive semantic analysis of construction contracts. NLP and machine learning algorithms were applied for the purposes of the automatic classification of contract clauses according to departmental relevance and accurately predicting the presence of ambiguity in contract clauses. The SVM and DT algorithms demonstrated the best performance.

The classification studies were also carried out for analysis of RFIs, but these classification studies were mostly performed manually (Tilley et al.,1997; Morales et al., 2022). In their studies on the analysis of RFIs, the researchers manually classified RFI documents under three main headings according to type, cause, discipline.

The manual classification of texts is acknowledged as time-consuming and error-prone. The process of manually categorizing thousands of RFI documents inevitably prolongs the time required to finalize the analysis. Additionally, misclassifications significantly impact the accuracy and reliability of the analysis outcomes concerning RFI documents. Consequently, the adoption of an NLP-based approach to analyze RFI documents offers potential benefits, such as shorter analysis times and more precise analysis results.

Methodology

Python programming language, Natural Language Toolkit and Scikit-learn library were used for natural language processing and machine learning methods applied to RFI documents. The original dataset is the RFI documents extracted from the common data environment of a project. In order to improve the quality of the original dataset, RFI documents with a lot of noise and very short descriptions were excluded. The metadata to be automatically retrieved from the RFIs was the selected as discipline (i.e., architectural, electrical, mechanical, and structural). The dataset comprises 25 RFI documents for each discipline. In total, 100 RFI text documents were analyzed. The RFI documents retrieved in .pdf format and each RFI document was turned into .txt format.

Implementation workflow for automatic extraction of RFI discipline metadata is shown in Figure 1. It has three main steps: Label assignment, NLP and machine learning. In the first step, each RFI document was labeled in terms of discipline. In the second step, NLP techniques were applied. NLP techniques were performed using the Natural Language Toolkit (NLTK) in Python. These blocks, also known as modules in NLTK, aid in tasks such as tokenization, stemming, lemmatization, and data classification. First, RFI documents were tokenized. Tokenization involves breaking a sequence of text into smaller units, referred to as tokens. These tokens can vary in size, ranging from individual characters to entire words

(Manning et. al., 2009). Second, punctuation marks and stop words were removed from each RFI document. Punctuation removal is often performed to diminish the dimensionality of the data and to eliminate elements that may not carry significant semantic meaning for certain NLP tasks. Thirdly, stop word removal was performed. It is a common preprocessing step in NLP where frequently occurring words, known as stop words, are removed from a text corpus. These words are often common (i.e., 'the', 'a', and 'in') and do not contribute significantly to the understanding of the content. Finally, RFI text data was converted into vectors via vectorization process. Vectorization techniques transform text data into a form that a computer can make sense of. In this study, Bag of Words (BOW) vectorization technique, one of the mostly used vectorization techniques (Qader et al., 2019), was used. BoW was utilized to represent text based on the number of word occurences, using a fixed-length vector created from a vocabulary. The BoW technique was implemented using the Scikit Learn Python library.

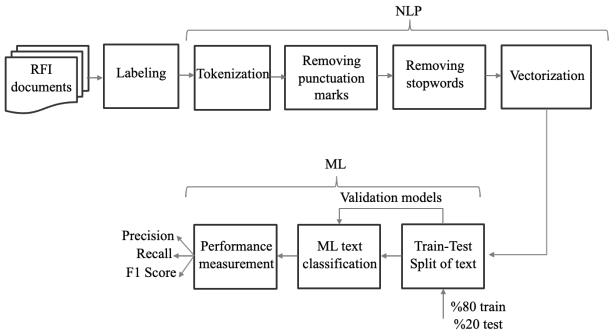

Figure 1: Implementation workflow for automatic extraction of RFI discipline metadata

In the third step, machine learning algorithms were applied. The automatic extraction of metadata from RFI documents is a classification problem. Therefore, supervised machine learning models need to be trained, and two classification algorithms, Naïve Bayes and K-Nearest Neighbor, were selected. Bayes' rule is used to calculate the probabilities of the classes. The NB algorithm is based on Bayesian Decision Theory (Alpaydin, 2014) and the NB model is the most widely used Bayesian network model in machine learning. The term "naïve" is used because the NB algorithm is based on the assumption that attributes are conditionally independent of each other (Russell and Norvig, 2010). It is extensively used in text categorization tasks, including document classification and spam e-mail detection.

The K-Nearest Neighbor classifier categorizes the input into the class with the highest frequency among the k neighbors of the input. Each neighbor holds an equivalent vote, and the class with the highest count of votes among the k neighbors is chosen. Ties are resolved arbitrarily or through a weighted vote. Typically, k is chosen as an odd number to reduce ties, especially when confusion occurs between two adjacent classes (Alpaydin, 2014).

First, the data was split into 80/20, where the first percentage represented the proportion of data allocated to the training set, and the second percentage represented the testing set. Validation was performed before testing the model within training set. The selected algorithms are trained to automatically extract RFI discipline metadata. Then the KNN and NB algorithms were trained and tested by using the output of the vectorization process.

Traditionally, two metrics are commonly used for measuring the performance of machine learning models: Precision and Recall. Precision measures the proportion of truly relevant documents in the result test. Recall measures the ratio of all relevant documents in the corpus that are included in the result set. It is possible to achieve a balance between precision and recall by adjusting the size of the returned result set. The F1-Score is the harmonic mean of sensitivity and recall (Russell and Norvig, 2010).

Precision, recall and F1Score values are calculated by substituting the parameters shown in Table 1 into the equations (1) to (3).

$$Precision = \frac{TP}{(TP+FP)} \quad (1)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

$$F1Score = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \quad (3)$$

Table 1: True and false positives and negatives

	Relevant	**Irrelevant**
Retrieved	True Positives (TP)	False Positives (FP)
Not retrieved	False Negatives (FN)	True Negatives (TN)

Findings

In the BoW vectorization process, word counts were performed for each discipline and they are used as input when applying the BoW technique. Twenty most common words related to the architectural discipline were determined and used in the vectorization process. The first five most repeated words out of the 20 most common words are given in Figure 2.

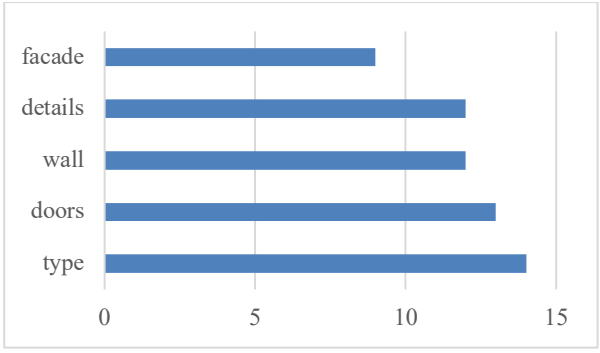

Figure 2: First five most common words in RFI documents related to the discipline of architecture

Similar to the architecture discipline, twenty most common words related to structural, mechanical and electrical disciplines were identified. The first five most repeated words out of the 20 most common are reported in Figure 3 to 5 for structural, electrical, and mechanical disciplines respectively.

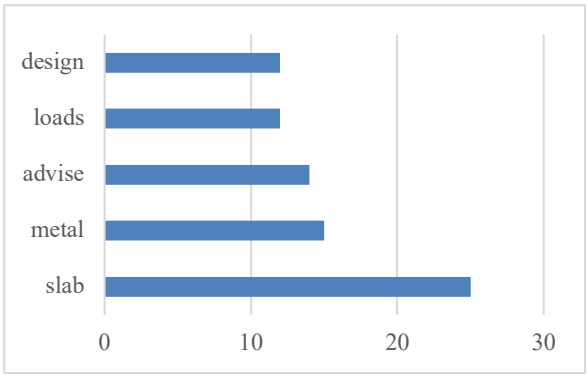

Figure 3: First five most common words in RFI documents related to structural discipline

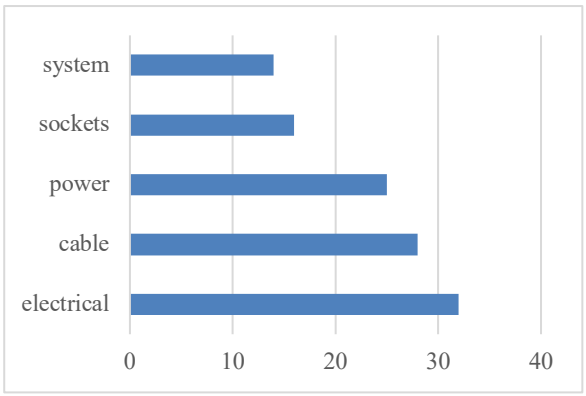

Figure 4: First five most common words in RFI documents related to electrical discipline

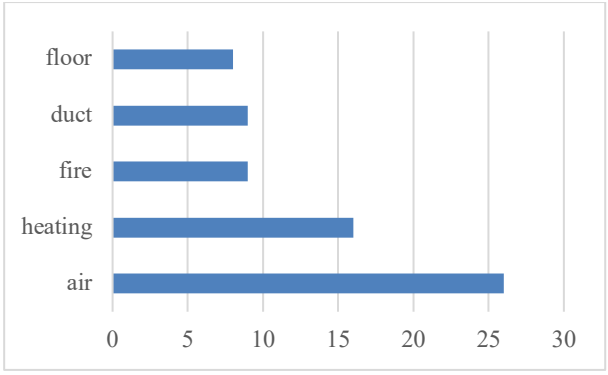

Figure 5: Some of the most common words in RFI documents related to mechanical discipline

After the vectorization process, the NB and KNN models were trained. %80 percent of the data was used for training and %20 percent for testing. The content of the eighty percent training set was changed ten times and validation work was carried before the test phase. At different iterations, the performance fluctuation of the trained algorithm was checked. The performance result of the trained NB algorithm is shown in Figure 6. After ten iterations, the Average Precision, Recall, and F1 Score values are approximately 85%, 82% and 82% respectively. The NB algorithm proves to be highly successful in automatically extracting the discipline metadata of RFIs.

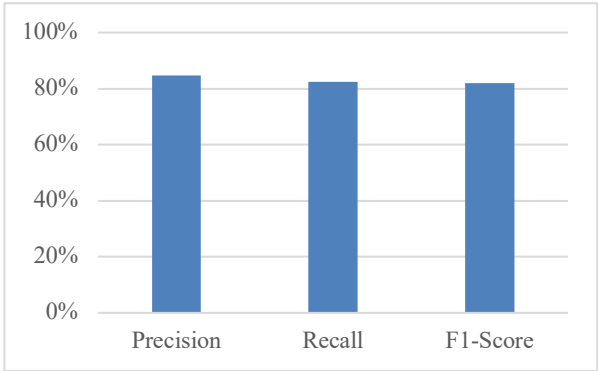

Figure 6: Performance results of the NB algorithm

The tuning of the hyperparameter k corresponding to the number of nearest neighbors is important for enhancing the performance of the KNN algorithm.

The performance results for different k numbers are determined and compared as shown in Figure 7. The results illustrated that KNN demonstrated the best performance when k was set to 5. Precision, Recall and F1 Score values are approximately %75, %60 and %62, respectively.

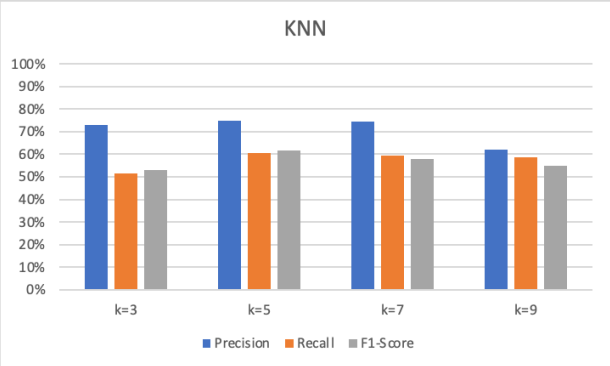

Figure 7: Performance results of KNN algorithm for different k values

Comparative performance results of NB and KNN algorithms are given in Figure 8. NB algorithm showed much better results than KNN algorithm according to precision, recall and F1-Score criteria. Overall, the findings have illustrated that one of the metadata of RFIs, which is discipline, can be successfully extracted automatically.

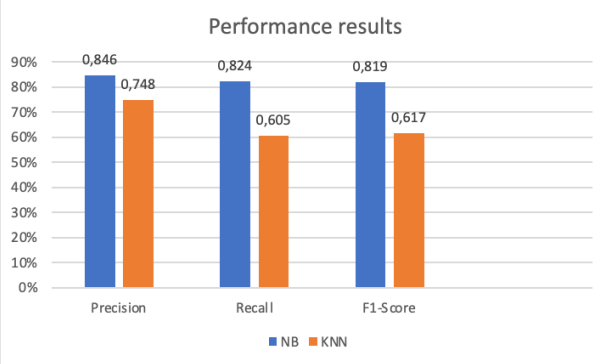

Figure 8: Comparative performance results of the algorithms

Conclusion

In this study, an NLP and machine learning-based model is developed to automatically extract the discipline, which is one of the metadata of RFIs. The performances of the two machine learning models, NB and KNN algorithms, are compared. The average precision of the NB algorithm is around 85%, while for the KNN algorithm, it is approximately 75%. In the case of KNN, it was demonstrated that the KNN algorithm gave better results for k=5.

The model proposed in this study has the potential to prevent time losses during the manual entry of metadata for RFIs in the common data environments, and it can contribute to a reduced number of incorrect entries.

An artificial intelligence based RFI management system can be developed by integrating NLP and machine learning models into the system. As a result, RFI management systems will be more effectively used. The utilization of emerging technologies and data-driven analytics can be harnessed to augment the RFI analysis process, leading to improvements in overall project efficiency (Afzal et al., 2023).

In future studies, we plan to conduct a more comprehensive analysis by expanding the dataset and diversifying the algorithms. In addition to training shallow machine learning models, it may be beneficial to explore the development of deep learning models.

References

Afzal, M., Wong, J.K.W, Fini, A.A.F. (2023) Unlocking insights: analysing construction issues in request for information (RFI) documents with text mining and visualisation. IEEE 19th International Conference on Automation Science and Engineering (CASE).

Aibinu, A.A., Carter, S., Francis, V., Vaz-Serra, P. (2019) Request for information frequency and their turnaround time in construction projects. Built Environment Project and Asset Management, Vol. 10, No.1, pp. 1-15.

Alpaydin, E. (2014) Introduction to Machine Learning, Third Edition. London, England, The MIT Press.

Bhat, A.C. (2015) Data visualization of requests for information to support construction decision-making (Master Thesis). The University of British Columbia.

Candas, A.B. (2022) Multipurpose semantic analysis of construction text documentation. A Thesis Submitted to the Graduate School of Natural and Applied Sciences of Middle East Technical University. Ankara.

Das, M., Tao, X., Cheng, J.C.P. (2020) A secure and distributed construction document management system using blockchain. Proceedings of the 18th International Conference on Computing in Civil and Building Engineering (ICCCBE).

Deloose, A., Gysels, G., Baets, B.D., Verwaeren, J. (2023) Combining natural language processing and multidimensional classifiers to predict and correct CMMS metadata. Computers in Industry, 145, 103830.

Eken, G. (2022) Using natural language processing for automated construction contract review during risk assessment at the bidding stage. A Thesis Submitted to the Graduate School of Natural and Applied Sciences of Middle East Technical University. Ankara.

Goh, Y.M., Ubeynarayana, C.U. (2017) Construction accident narrative classification: an evaluation of text mining techniques. Accident Analysis and Prevention, 108, 122-130.

Han, H., Giles, C.L., Manavoğlu, E., Zha, H. (2003) Automatic document metadata extraction using support vector machines. Proceedings of the Joint Conference on Digital Libraries.

Hanna, A. S., Tadt, E.J., Whited, G.C. (2012) Request for information: benchmarks and metrics for major highway projects. Journal of Construction Engineering and Management, 138(2), 1347-1352.

Kelly, D. and Llozor, D.B. (2020) Performance outcome assessment of the integrated project delivery (IPD) method for commercial construction projects in USA. International Journal of Construction Management, 1-9.

Manning, C.D., Raghavan, P., Schütze, H. (2009) An information to information retrieval. Cambridge University Press, Cambridge, England .

Morales, F., Herrera, R.F., Rivera, F.M.-L., Atencio, E., Nunez, M. (2022) Potential application of BIM in RFI in building projects. Buildings, 12, 145.

Paik, W., Yılmazel, S., Brown, E., Poulin, M., Dubon, S., Amice, C. (2001) Applying natural language processing (NLP) based metadata extraction to automatically acquire user preferences. K-CAP'01, October 22-23, Victoria, British Columbia, Canada.

Qader, W.A., Ameen, M.M., Ahmed, B.I. (2019) An overview of bag of words; importance, implementation, applications, and challenges. International Engineering Conference.

Philips-Ryder, M., Zuo, J., Jin, X.H. (2012) Evaluating document quality in construction projects – subcontractors' perspective. International Journal of Construction Management 13 (3) 77– 806 94.

Russell, S. and Norvig, P. (2010) Artificial Intelligence: A Modern Approach, Third Edition. Upper Saddle River, New Jersey 07458, Pearson.

Shim, E., Carter, B., Kim, S. (2016) Request for information (RFI) management: a Case Study. 52nd ASC Annual International Conference Proceedings.

Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., Raynal, C. (2016) Natural language processing for aviation safety reports: from classification to interactive analysis. Comput. Ind. 78, 80–95.

Tilley, P.A.; Wyatt, A.; Mohamed, S. Indicators of design and documentation deficiency. In Proceedings of the IGLC-5, Fifth Annual Conference of the International Group for Lean Construction, Gold Coast, Australia, 16–17 July 1997; pp. 137–148.

Tixier, A.J.P., Hallowell, M.R., Rajagopalan, B., Bowman, D. (2016) Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports. Automation in Construction, 62, 45-56.

Tixier, A.J.P., Hallowell, M.R., Rajagopalan, B., Bowman, D. (2016) Application of machine learning to construction injury prediction. Automation in Construction, 69, 102- 114.

Uz-Zaman, K.A., Cholette, M.E., Ma, L., Karim, A. (2017) Extracting failure time data from industrial maintenance records using text mining. Adv. Eng. Inform. 33, 388–396.

Valdez, J., Rueschman, M., Kim, M., Redline, S., Sahoo, S.S. (2016) An ontology- enabled natural language processing pipeline for provenance metadata extraction from biomedical text. OTM Conferences, LNCS 10033, 699-708.

Yılmaz, İ.C., Dikbaş, A. (2013) Türk kamu inşaat projelerinde yaşanan uyuşmazlıklara yönelik bir veri madenciliğI yaklaşımı. Online Academic Journal of Information Technology, Cilt: 4, Sayı:13, doi: 10. 5824/1309-1581.2013.4.005.x.

Yılmaz, İ.C. (2013) İnşaat sözleşmelerinde hak talebi yönetimi: kamu projeleri için öneri model (Doktora Tezi). İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü.

Yılmazel, O., Finneran, C.M., Liddy, E.D. (2004) MetaExtract: an NLP system to automatically assign metadata. Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04).

Zhang, T., Bhatia, A., Pandya, D., Sahinidis, N.V., Cao, Y., Flores-Cerrillo, J. (2020) Industrial text analytics for reliability with derivative-free optimization. Comput. Chem. Eng. 135, 106763.