ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

Automated Construction Activities Extraction from Accident Reports Using Deep Neural Network and Natural Language Processing Techniques

Quan Do[1]*, Tuyen Le[2], Chau Le[3]

[1] *The Glenn Civil Engineering Department, Clemson University, 131 Lowry Hall, Clemson, SC 29634, USA,* E-mail address: qdo@clemson.edu
[2] *The Glenn Civil Engineering Department, Clemson University, 316 Lowry Hall, Clemson, SC 29634, USA,* E-mail address: tuyenl@clemson.edu
[3] *Department of Civil, Construction and Environmental Engineering, North Dakota State University, 1410 14th Avenue North, Fargo, ND 58102, USA,* E-mail address: chau.le@ndsu.edu

Abstract: Construction is among the most dangerous industries with numerous accidents occurring at job sites. Following an accident, an investigation report is issued, containing all of the specifics. Analyzing the text information in construction accident reports can help enhance our understanding of historical data and be utilized for accident prevention. However, the conventional method requires a significant amount of time and effort to read and identify crucial information. The previous studies primarily focused on analyzing related objects and causes of accidents rather than the construction activities. This study aims to extract construction activities taken by workers associated with accidents by presenting an automated framework that adopts a deep learning-based approach and natural language processing (NLP) techniques to automatically classify sentences obtained from previous construction accident reports into predefined categories, namely TRADE (i.e., a construction activity before an accident), EVENT (i.e., an accident), and CONSEQUENCE (i.e., the outcome of an accident). The classification model was developed using Convolutional Neural Network (CNN) showed a robust accuracy of 88.7%, indicating that the proposed model is capable of investigating the occurrence of accidents with minimal manual involvement and sophisticated engineering. Also, this study is expected to support safety assessments and build risk management systems.

Key words: Natural language processing, Deep learning, Convolutional Neural Network, Sentence classification, Accident reports

2. INTRODUCTION

The International Labor Organization (ILO) estimates that more than 1.9 million people die from occupational accidents and work-related illnesses each year, and 90 million people have disability-adjusted life years (DALYs) [1]. The construction field caused one of six deaths during the period, playing a major role in accidents that caused health issues, lost time, and financial loss [2]. Although some accidents happen in unexpected and unusual ways, most previous accidents are identical in certain aspects. It is critical to investigate previous incidents and understand the reasons to avoid similar incidents and increase workplace health and safety. Given the importance of accident reports, greater emphasis has been placed recently on guaranteeing the quality of data gathering and report administration.

The information in accident reports is currently underused due to the difficult extraction from unstructured accident reports. The initial step in analyzing construction accident records effectively is information classification before performing additional analytics. It requires a significant amount of time and resources to read the text data and ensure that the categorization results are consistent due to human abilities and ever-increasing volumes of data. As a result, it is critical to develop the approach of automatic classification of text data in accident reports. Recently, the-state-of-the-art machine learning algorithms resulted in several considerable efforts to develop classification models by researchers. Tixier et al. [3] applied machine learning algorithms to predict the energy type, injury type, and the injured body part. Chokor et al. [4] presented an unsupervised approach to classify types of injury reports. Those models could not extract the information regarding the construction activities associated with accidents.

The goal of this study is to extract construction activities taken by workers associated with accidents. This study adopted convolutional neural network (CNN) and natural language processing (NLP) techniques for automatically classifying sentences of accident reports using the Occupational Safety and Health Administration (OSHA) data into three categories associated with the type of information. Those categories are Trade (i.e., a construction activity before an accident), Event (i.e., an accident), and Consequence (i.e., the outcome of an accident). The results can provide useful insights to support safety assessments and build risk management systems.

3. BACKGROUND

3.1. NLP and Sentence Classification

NLP is a subfield of Artificial Intelligence and Linguistics that concentrates on giving computers the ability to understand and analyze the natural language, such as text or speech [5]. Natural language is the ordinary language in which people speak or write for general communication purposes. NLP is applied to a wide range of tasks such as translation, spam detection, information extraction, summarization, and question answering. In many areas of NLP, text classification is an important task, which is used to classify free-text documents into predefined categories and can be used in a variety of industries, such as medical [6], financial [7], and social analysis [8]. In construction, text classification has been frequently employed to solve a variety of issues such as supporting construction field inspection [9], compliance checking [10], and enhancing management efficiency [11].

Sentence classification is a specialized form of text classification that categorizes text sentences into distinct groups based on their grammatical and semantic structure. The amount of text data has increased dramatically throughout the last few decades, which has both facilitated and complicated the task of sentence classification, necessitating the development of more robust and scalable possible solutions. With the recent advances in Machine Learning, our computers can now analyze, comprehend, and extract information from extensive complex sentences [12].

3.2. CNN

CNN was initially invented for computer vision and has been shown as a powerful tool in image analytics [13]. CNN models were later shown to be effective for NLP, with impressive results in sentence modeling [14], semantic parsing [15], and other NLP tasks [16], [17]. With the application of word embedding, each sentence can be formed as a matrix; this mapping approach has motivated researchers to use CNN for sentence classification [18]. CNN models have demonstrated outstanding performance on sentence classification problems, owing to the ability to extract local features by using convolution layers and accumulate global information by constructing hierarchical structures [19]. Kim [20] presented the earliest effort, which examined the capability of CNN on sentence classification and achieved excellent results. Several successful studies [21], [22] have proved the superior performance of CNN in sentence and text classification tasks.

3.3. Related Studies

Recently, the the-state-of-the-art of machine learning algorithms resulted in several considerable efforts to develop classification models by researchers [23], [24]; these studies classify the cause of construction accidents and identify frequent objects causing accidents. Tixier et al. [3] proposed an automated model based on hand-coded rules and keywords (lexicon) to classify construction incident narratives from 2201 unstructured injury reports. Researchers used lexicon to reduce the large range of keywords and achieved an accuracy of 95%. Chokor et al. [4] conducted a K-means clustering unsupervised approach to classify construction injury reports. Four types of accident causes were identified, including fall, struck by objects, electrocutions, and trench collapse. In other research, Goh and Ubeynarayana [23] used Naïve Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) algorithms for classifying construction accident. These algorithms classified 1000 construction reports obtained from OSHA into 11 labels of accident causes. In 2020, Zhong et al. [25] introduced the application of Convolutional Neural Network to automatically classify causes of construction accidents and the Latent Dirichlet Allocation model to examine the interdependency between causal variables. Those models only extracted the causes and related objects and could not extract the information regarding the construction activities associated with accidents. This study is to address the research gap in previous studies; we propose an automated model that classifies sentences and helps extract construction activities taken by workers associated with accidents. The results can provide useful insights to support safety assessments and build risk management systems.

4. METHODOLOGY

This study presents classification models that classify extracted sentences from accident reports into several predefined categories, namely Trade, Event, and Consequence. Trade label refers to construction activities before an accident happens like Roofing, Excavation, and Carpentry; whereas Event label refers to an accident happening such as Fall, Crush, or Electric Shock. Consequence label indicates the outcome of the accident that has occurred, such as Fracture and Fatality. To begin with, the dataset was developed for training and evaluating the model. In the following step, several techniques were used to preprocess the dataset. Afterward, the proposed methodology employed the Glove model to present text as numerical values. Finally, the CNN-based approach performed sentence classification before model evaluation was implemented.

4.1. Dataset Preparation

The accident reports are available for download for free at the OSHA website [26]. In this research, the accident reports containing a detailed account of the accidents on construction sites were selected and stored in a Microsoft Excel file. A sample size of 380 accident reports was chosen, and each report was split into separate sentences. As a result, the raw dataset of 2,158 sentences was extracted. The authors manually annotated these sentences into predefined labels, namely Trade, Event, and Consequence. The labeled datasets later were reviewed by another researcher in the same domain. Since there were no inconsistencies from the reviewer, the labels were formalized for the statements.

Accident reports	Statements	Label
On December 4 2013 Employee #1 a carpenter employed by Valley Trinity Construction Co. Inc. was engaged in interior carpentry work at a commercial building. He fell from a ceiling joist a fall height of approximately 6 feet. Emergency services were called and Employee #1 was transported to a hospital where he was admitted and treated for bruising/abrasions to his back and neck.	On December 4 2013 Employee #1 a carpenter employed by Valley Trinity Construction Co. Inc. was engaged in interior carpentry work at a commercial building.	Trade
	He fell from a ceiling joist a fall height of approximately 6 feet.	Event
	Emergency services were called and Employee #1 was transported to a hospital where he was admitted and treated for bruising/abrasions to his back and neck.	Consequence

Figure 1. Sample accident reports and labeled statements

4.2. Data Preprocessing

Several NLP techniques were used to preprocess the developed datasets, including lowercasing and punctuation removal, tokenization, and lemmatization. Lowercasing fundamentally transformed text to lowercase to guarantee that terms with similar meanings and punctuation marks were already eliminated as they were useless for sentence classification tasks. Tokenization is the process that separates a sentence into small pieces called tokens, which might be a single word, number, punctuation, or blank space [27]. Lemmatization refers to reducing various forms of a word to its root form. For example, the words "performing," "performs," and "performed" were converted to "perform." Lemmatization improves the model's performance and computing effectiveness since it utilizes only a form to represent distinct grammatical versions of a word [28].

4.3. Text Representation

Since deep neural networks are not able to deal directly with words, the word embedding process is carried out to convert text to numerical representation that the algorithm can recognize and handle. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. The GloVe model learns by looking at each pair of words that might co-occur in the corpus and constructing a co-occurrence matrix. GloVe uses an objective function to train word vectors from the co-occurrence matrix. On the Glove website, several pre-trained word vector databases are available for the public. The use of pre-trained word embeddings is advantageous as they trained on millions of words instead of training new word embeddings from scratch.

4.4. CNN-based classification

In this study, the architecture of the CNN-based sentences classification model used for the training is presented in Figure 2. There are four layers in total, including the word embedding layer, the convolution layer, the max-pooling layer, and the fully connected layer. The Word embedding layer employs the embedding matrix to convert the sequence of words in the input sentence to the matrix that plays as input for the subsequent convolution layer. Next, the convolution layer uses multiple filters to extract the features from the input sentence presentation matrix. The max-pooling layer is in charge of minimizing the spatial size of the complicated features obtained from the output of the convolution layer. The fully connected layer consists of hidden layers, and the output layer is trained to classify sentences automatically. The number of predefined categories determines the number of neurons in the output layer.

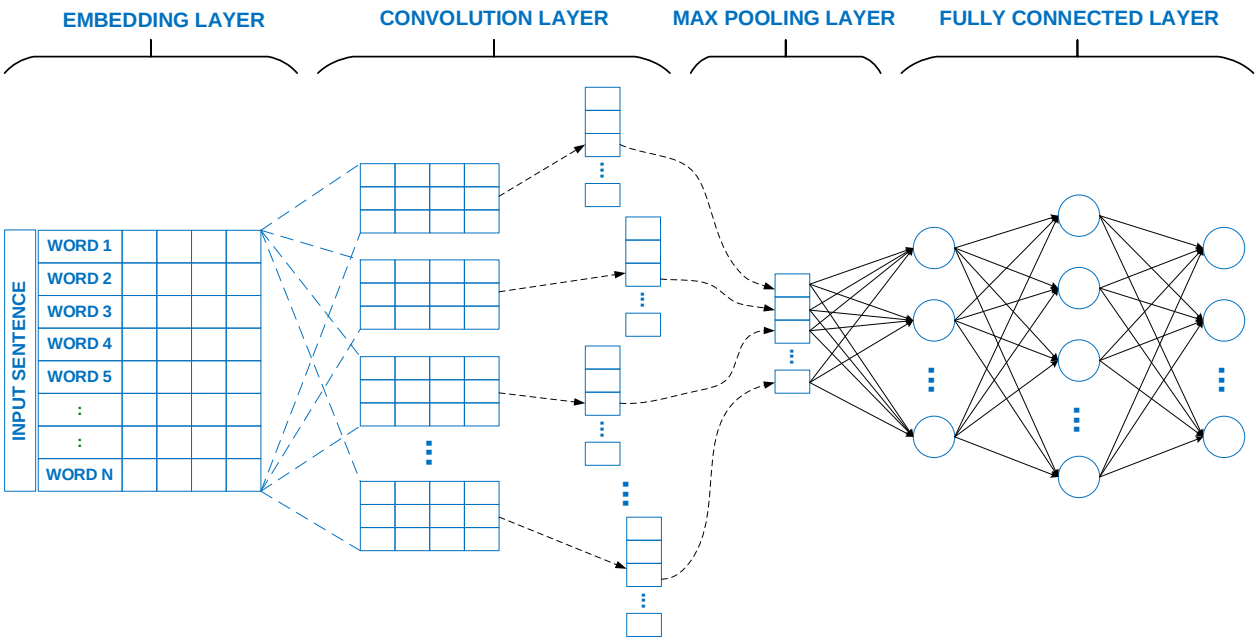

Figure 2. Architecture of the CNN-based sentences classification model

The parameters of the CNN model were tuned to achieve optimal performance. The grid of parameters was defined. The convolution layer had the numbers filters of 32, 64, and 128, besides the filter sizes of 3,5 and 7. A dropout technique with a probability of 0.5 was adopted to avoid overfitting, and a 5-fold cross-validation was applied. The CNN model had the following hyperparameters making the optimal result: number of epochs = 100; batch size = 100; Activation function of hidden layer was ReLU; Activation function of output layer was Softmax; Loss function was Sparse categorical cross entropy; Optimizer was Adam; number of filters = 128; filter size = 3.

4.5. Model Evaluation

The dataset was randomly split into 80% and 20% for the training set and testing set, respectively. Classification evaluation is to evaluate the performance of the model on the testing set. Performance metrics mainly used are accuracy, precision, recall, and F1-score. F1-score is the robust measure widely used considering both precision and recall.

5. RESULTS AND DISCUSSION

The model had an accuracy of 88.7% and the highest f1-score of 0.92. The performance of the classification model in classifying sentences is a satisfactory result (as depicted in Table 1). It can be seen that the performance metrics of Trade and Event are relatively similar, with the average precision, recall, and f1-score being 0.91, 0.92, and 0.91, respectively. In terms of Consequence, its performance metrics are 0.82 and 0.76, respectively, for precision and recall, which are both lower than those of Trade and Event. In comparison with Trade and Event, Consequence has a poorer f1-score of 0.79, yet it still represents an excellent value.

Table 1. Performance metrics of the classification model

Label	precision	recall	f1-score	support
Trade	0.89	0.92	0.90	278
Event	0.92	0.92	0.92	154
Consequence	0.82	0.76	0.79	108

In general, the classification model achieved promising performance metrics. A number of incorrect predictions are primarily due to inaccurate natural language, which made the model confusing between labels. For example, the sentence "He was flown to MD" has the actual label of Event; however, it was predicted as Consequence label. This example could show the most popular reasons behind incorrect predictions are unclear semantic relationships, short text, and abbreviation due to free text in construction accident reports. Several statements describe the outcome of an accident; the records also include additional evidence and information from later investigations. These statements are pretty close to sentences with Trade labels, resulting in incorrect predictions. For example, the sentence "The coworker contacted two other workers at the site to assist with the rescue of Employee #1 and the workers performed CPR on Employee #1 until an ambulance arrived on scene" has the actual label of Consequence; however, it was predicted as Trade since the information provided similar to trade sentences. The authors also propose that labels with low f1-scores undergo further manual reviews.

6. CONCLUSION

Construction accident reports are informative available documentation, and the process of analyzing them may allow gaining a vital understanding of past occurrences to prevent future reoccurring incidents. Besides, manual investigation of accident reports is time-consuming; thus, automatic implementation is expected to save time and be capable of serving for further analysis. This study aims to extract construction activities taken by workers from accident reports. To achieve this goal, we propose an automated framework that adopts a deep learning-based approach and natural language processing (NLP) techniques to automatically classify sentences obtained from previous construction accident reports into predefined categories, namely Trade, Event, and Consequence. This information can help enhance our understanding of historical data and be utilized for accident prevention.

The sentence classification model was trained on a dataset of 2,158 sentences extracted from OSHA construction accident reports. The model had an accuracy of 88.7% and the highest f1-score of 0.92, which were sufficient for satisfactory sentence classification accident reports.

This research offers substantial contributions to the body of knowledge. This is the first study that employed the deep neural network and NLP to classify sentences in accident reports into informative categories. A reliable classification model was developed that can be further reconstructed to exploit various information from each meaningful sentence extracted from construction accident records. In the realm of practice, construction firms can adopt this automated model instead of manual implementation that can save time and resources in analyzing and extracting information from accident reports. Furthermore, this study is expected to help the safety assessments and build risk management systems to prevent financial loss and catastrophe.

This study showed favorable results; however, one significant limitation is the small size of the dataset. As supervised machine learning, classification using CNN is trained on the labeled dataset. The more examples are learned, the more distinctive features are extracted in the convolution layer, which improves classification accuracy.

REFERENCES

[1] International Labor Organization (ILO), "Safety and health at work." http://www.ilo.org/global/topics/safety-and-health-at-work/lang--en/index.html (accessed Dec. 04, 2021).

[2] F. Zhang, H. Fleyeh, X. Wang, and M. Lu, "Construction site accident analysis using text mining and natural language processing techniques," *Automation in Construction*, vol. 99, pp. 238–248, Mar. 2019, doi: 10.1016/j.autcon.2018.12.016.

[3] A. J. P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports," *Automation in Construction*, vol. 62, pp. 45–56, Feb. 2016, doi: 10.1016/j.autcon.2015.11.001.

[4] A. Chokor, H. Naganathan, W. K. Chong, and M. el Asmar, "Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning," in *Procedia Engineering*, 2016, vol. 145, pp. 1588–1593. doi: 10.1016/j.proeng.2016.04.200.

[5] C. Manning and H. Schutze, *Foundations of statistical natural language processing*. MIT press, 1999.

[6] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classification using convolutional neural networks," in *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, IOS Press, 2017, pp. 246–250.

[7] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 2, pp. 1–19, 2009.

[8] P. Jotikabukkana, V. Sornlertlamvanich, O. Manabu, and C. Haruechaiyasak, "Effectiveness of social media text classification by utilizing the online news category," in *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 2015, pp. 1–5.

[9] N. W. Chi, K. Y. Lin, N. El-Gohary, and S. H. Hsieh, "Evaluating the strength of text classification categories for supporting construction field inspection," *Automation in Construction*, vol. 64, pp. 78–88, Apr. 2016, doi: 10.1016/j.autcon.2016.01.001.

[10] D. M. Salama and N. M. El-Gohary, "Semantic Text Classification for Supporting Automated Compliance Checking in Construction," *Journal of Computing in Civil Engineering*, vol. 30, no. 1, p. 04014106, Jan. 2016, doi: 10.1061/(asce)cp.1943-5487.0000301.

[11] D. Tian, M. Li, J. Shi, Y. Shen, and S. Han, "On-site text classification and knowledge mining for large-scale projects construction by integrated intelligent approach," *Advanced Engineering Informatics*, vol. 49, Aug. 2021, doi: 10.1016/j.aei.2021.101355.

[12] X. Jiang, B. Zhang, Y. Ye, and Z. Liu, "A hierarchical model with recurrent convolutional neural networks for sequential sentence classification," in *CCF International Conference on Natural Language Processing and Chinese Computing*, 2019, pp. 78–89.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[14] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A Convolutional Neural Network for Modelling Sentences." 2014.

[15] W. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 643–648.

[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.
[17] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proceedings of the 23rd international conference on world wide web*, 2014, pp. 373–374.
[18] L. Zhining, G. Xiaozhuo, Z. Quan, and X. Taizhong, "Combining Statistics-Based and CNN-Based Information for Sentence Classification; Combining Statistics-Based and CNN-Based Information for Sentence Classification," *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2016, doi: 10.1109/ICTAI.2016.153.
[19] J. Shin, Y. Kim, S. Yoon, and K. Jung, "Contextual-CNN: A Novel Architecture Capturing Unified Meaning for Sentence Classification," in *Proceedings - 2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018*, May 2018, pp. 491–494. doi: 10.1109/BigComp.2018.00079.
[20] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.
[21] F. Wei, H. Qin, S. Ye, and H. Zhao, "Empirical Study of Deep Learning for Text Classification in Legal Document Review," in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, Jan. 2019, pp. 3317–3320. doi: 10.1109/BigData.2018.8622157.
[22] J. Hoffimann, Y. Mao, A. Wesley, and A. Taylor, "Sequence Mining and Pattern Analysis in Drilling Reports with Deep Natural Language Processing," Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.01476
[23] Y. M. Goh and C. U. Ubeynarayana, "Construction accident narrative classification: An evaluation of text mining techniques," *Accident Analysis and Prevention*, vol. 108, pp. 122–130, Nov. 2017, doi: 10.1016/j.aap.2017.08.026.
[24] M. Y. Cheng, D. Kusoemo, and R. A. Gosno, "Text mining-based construction site accident classification using hybrid supervised machine learning," *Automation in Construction*, vol. 118, Oct. 2020, doi: 10.1016/j.autcon.2020.103265.
[25] B. Zhong, X. Pan, P. E. D. Love, L. Ding, and W. Fang, "Deep learning and network analysis: Classifying and visualizing accident narratives in construction," *Automation in Construction*, vol. 113, May 2020, doi: 10.1016/j.autcon.2020.103089.
[26] Occupational Safety and Health Administration (OSHA), "Fatality and Catastrophe Investigation Summaries." https://www.osha.gov/pls/imis/accidentsearch.html (accessed Dec. 11, 2021).
[27] G. Grefenstette and P. Tapanainen, "What is a word, what is a sentence?: problems of Tokenisation," 1994.
[28] V. Balakrishnan and L.-Y. Ethel, "Stemming and Lemmatization: A Comparison of Retrieval Performances," *Lecture Notes on Software Engineering*, vol. 2, pp. 262–267, Dec. 2014, doi: 10.7763/LNSE.2014.V2.134.