Australasian Transport Research Forum 2021 Proceedings
8-10 December, Brisbane, Australia
Publication website: http://www.atrf.info

A conceptual approach towards the evaluation of the vulnerability of urban railway network infrastructure by analysing railway accident reports

Wei-Ting Hong[1], Geoffrey Clifton[1], John D Nelson[1]

[1] Institute of Transport and Logistics Studies, The University of Sydney

Email for correspondence: wei-ting.hong@sydney.edu.au

1. Introduction

Railway transportation plays an essential role in the functioning of society, economy, and environment. As part of a complex multi-modal urban transport network, railway transportation serves as the backbone by handling a massive volume of passengers combined with efficient land use and accessibility. However, the urban transport network can be degraded significantly once the railway system experiences disruption, resulting in significant impact. The triggers of disruptive events are various (e.g. human error, infrastructure failure, natural events) and play an important role in the railway safety context. Unfortunately, our knowledge of identifying hazards and estimating the probability that each hazard triggers an incident are primarily based on minimal data in the literature, resulting in most studies tending to focus on the impact management and resources allocation based on assuming that the disruption has happened (Dehghani et al., 2014; Kim & Yeo, 2017) rather than concentrating on the nature of the hazards themselves. Additionally, a main challenge in railway safety research is to obtain analysable data. Part of the reason is that most railway accidents are described or recorded by texts instead of numbers. This means numerical data relating to railway accidents is scarce, but textual data is rich. On the other hand, although individual railway accidents can be analysed by experts of each authority adequately, understanding the nature of hazards from all railway accidents historically and across countries is still a challenging task even for railway safety experts. Hence, introducing automation to help us understand the nature of railway accidents comprehensively has become a critical option to improve railway safety.

Therefore, the motivation of this paper is to evaluate and improve our understanding of the vulnerability of urban railway networks through introducing state-of-the-art techniques to analyse railway accident reports and extract critical insights to meet the needs of the railway industry. Specifically, the scope encompasses the interactions between hazards, rail accident investigation bodies, and the operators. These relationships are demonstrated in Figure 1. The hazards are the core components within an accident, and the specific combination of hazards will be the trigger of an actual accident (Rausand, 2013). This study analyses the output from the independent railway accident investigation bodies on the basis of interest in improving railway safety. The reports written by independent railway accident investigation bodies contain a large amount of information, such as railway operators involved, the sequences of the events, the cause-effect analysis, and the recommendations. This content enables us to not only identify the underlying pattern of railway accidents, but also identify the way that hazards contribute to the vulnerability of railway networks.

The content of previous accidents is mostly recorded through text instead of numerically, hindering statistical analysis and resulting in the difficulty of extending horizontal knowledge in this context. Even in academia, techniques to solve such issues are manually demanding and time-consuming due to the availability and analysability of the data. To overcome this difficulty, Natural Language Processing (NLP) will be utilized to handle and classify the textual data, and an ontology will be introduced to describe the nature of the railway accidents and provide a solid framework to extract data from original reports written by railway accident investigation bodies from several English-speaking jurisdictions. Through decomposing the original reports in a logical way with an appropriately designed ontology, the comparison across different time and countries is applied, and the results can help urban railway network planners to reveal the underlying hazards and control the vulnerability of the urban railway network.

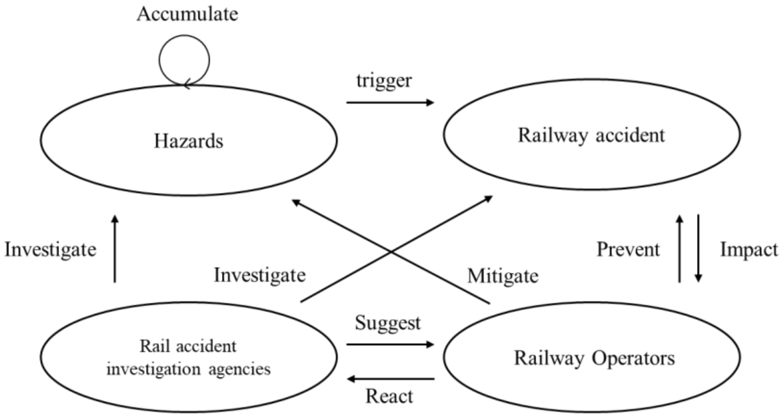

Figure 1: The proposed process of this study and the required resources.

2. Concise Literature Review

In the literature, two approaches are applied to analyse textual data as railway accident reports: manual analysis and Natural Language Processing (NLP). The former deconstructs the original reports to gain the probability of each event within an accident and manually highlights the critical elements based on a set of pre-defined rules. (see Kim & Yoon (2013) and Zhou & Lei (2018)) However, a labour-intensive analysis method has limitations. For instance, the number of available reports might be thousands or tens of thousands for a cross-nation analysis, and the cost would be unaffordable. Additionally, if we are going to propose a model with new variables that need to be abstracted from reports, all data must be read again.

On the other hand, the NLP allows researchers to develop insights from accidents and efficiently disseminate the information. The NLP is a technique that enables a computer to analyse textual data and generate a summary or horizontal conclusion through reading an enormous number of words in articles. In the context of railway accident report analysis, several document-level NLP models are designed to classify accident reports into relevant categories at the early stage, like human error, technology issues, and organization issues (Heidarysafa et al., 2019; Li et al., 2018).

However, most railway safety data is recorded through unstructured text but with solid description on the sequence of events and the insights. Accident reports usually contain a very complex sentence structure, and several connections would exist between sentences in the reports as well. The challenge of integrating data from multiple, unrelated sources into a unique framework for panel analysis results in limitations in exploring the knowledge in the context of railway safety (Katsumi & Fox, 2018).

3. Proposed Methodology

This study aims to identify the hazards in railway systems and understand how the railway safety agencies and the railway operators react to railway accidents. To achieve this goal, an analysis of the original official railway accident reports with state-of-the-art techniques is conducted to meet the academic and practical needs. The NLP approach has been widely considered as one of the best methods for analysing textual data due to the high performance, low time consumption, and rich flexibility of model design compared with other traditional approaches (Kulkarni &Shivananda, 2019; Young et al., 2018). Most importantly, unlike other traditional approaches such as rule-based approaches or web-crawlers, the NLP model is able to understand the real meaning behind the texts under a specific context with the help of machine learning algorithms. For example, traditional approaches can struggle to distinguish the difference between the word “train” used as a verb and a noun in the same article. However, the NLP model can easily accomplish that by extracting not only the features of that word, but the words that accompany it. Lastly, having an ontology as the framework allows us to map the knowledge of domain railway accidents extracted from the original data through NLP techniques.

The proposed process of this study is illustrated in Figure 2. For the process of retrieving data, a set of criteria is designed to select the countries whose railway accident reports are to be analysed. Apart from the structure, features and complexity of reports, the limitation of the methodology is also taken into account in the selection of countries. Once the countries are confirmed (in this case USA, UK, Australia, and Canada), a simple web crawler is designed to collect the official accident reports from the jurisdictions’ websites. The python package “*Beautiful Soup”* is applied for extracting documents from HTML (Hajba, 2018). We first search the official website containing all the accident reports to extract all document files in PDF format from the target website and its sub-website(s). Next, we filter out non-related documents based on their document file names and their contents. Only documents that have the sequence of the incidents, consequences of the accident, and recommendations section would be considered in our dataset.

Then, in the first step of NLP model development, the input data is pre-processed by lowercasing, lemmatizing, stop words removing, and tokenizing. Once the data is processed, we need to set the NLP objects and corresponding tasks to reach our research topic. The task Named Entity Recognition (NER) is selected for information extraction. The model will be able to identify the Interest of Entities (IOE) through analysing the characteristics of input data via machine learning algorithms. However, given the training data for railway accident reports is unavailable, the data will be annotated manually or via semi-supervised learning techniques.

After the tasks have been well set, the language model is built to solve the NLP tasks by appropriate NLP pipeline design. After the NLP pipeline is designed and well-evaluated, several informative contents will be extracted in a set of categories. To structure the outcomes from the NLP model, we construct the ontology to meet our interests. For the development of the ontology, the initial stage is to define the domain of interest and appropriate terminology. At this point, the upper-level ontology will be defined. Subsequently, the ontology will keep upgrading based on the output from the NLP model and discussion with experts from academia and industry. Finally, the process of structuring the ontology is completed after exhausting the data, and the completed ontology can now be applied for upcoming new data and further analysis.

Meanwhile, another NLP model is developed for the purpose of identifying whether railway accident investigation bodies learn from each other whilst coming up with recommendations. Hence, the recommendations made by different railway accident investigation bodies are retrieved individually and pre-processed. Next, the data is classified based on its features, and the distribution of data is demonstrated to reveal the difference. Finally, the differences will be highlighted and incorporated as a part our ontology.

In terms of evaluation, three prevalent measures, including precision, recall, and F-score (Li et al., 2020; Young et al., 2017), will be used to evaluate the performance of our model. We will manually annotate some data as gold standards (or correct labels). The model can be considered well-performing if most of its predictions meet the gold standards. The precision (p) is defined as $p = \frac{TP}{TP+FP}$, recall (r) is defined as $r = \frac{TP}{TP+FN}$, and F-score (F) is defined as $F = \frac{2pr}{p+r}$,where TP, FP, and FN represent true positive, false positive and true negative respectively.

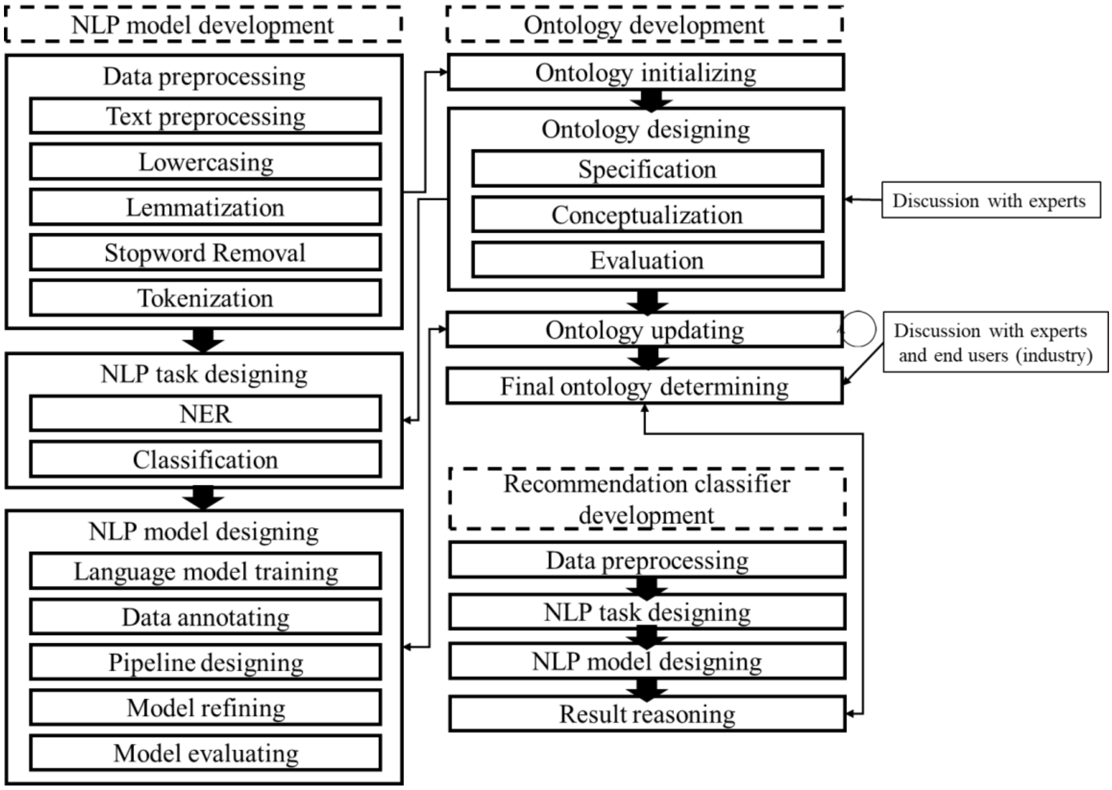

Figure 2: The proposed process of this study and the required resources.

4. Expected form of Results

The aim of this study is to design a framework for the industry to evaluate the vulnerability of the railway network and understand the nature of the railway accidents. The difficulty of retrieving textual data will be eliminated through the utilization of the NLP technique, and an exhaustive ontology will be generated via decomposing the original reports from several countries. Briefly, the following contributions are expected to be provided:

(1.) A domain ontology for the railway accidents domain containing the critical components causing the railway accidents and the relations between individuals within the railway system. The hazards for further improvement would be revealed.

(2.) An NLP model for describing the interface between original railway accident reports and the designed ontology. The critical components would be identified and allocated into the ontology automatically. Additionally, the model should be able to suggest updates to the structure of the ontology.

(3.) An NLP model for classifying the recommendations made by different railway investigation agencies. The categories can be defined by experts and the model per se, and a statistical analysis would be conducted to illustrate the difference between agencies from different countries.

(4.) Knowledge and experience mixing both practice and academic research in the context of railway safety through discussion on how to identify hazards in a railway system and the strategy of prevention.

However, there are some limitations. The result of this study cannot be used to predict new hazards after application of new technology. Additionally, the scope is restricted to only our selected countries and interested topics. Future research is recommended to concentrate on building ontologies for other topics, and the technique of data annotation for NLP model training is also a worthwhile further development.

5. Reference

Dehghani, M. S., Flintsch, G., &McNeil, S. (2014). Impact of road conditions and disruption uncertainties on network vulnerability. *Journal of Infrastructure Systems*, *20*(3). https://doi.org/10.1061/(ASCE)IS.1943-555X.0000205

Hajba, G. L. (2018). Website Scraping with Python. In *Website Scraping with Python*. Apress. https://doi.org/10.1007/978-1-4842-3925-4

Heidarysafa, M., Kowsari, K., Barnes, L., &Brown, D. (2019). Analysis of Railway Accidents' Narratives Using Deep Learning. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 1446–1453. https://doi.org/10.1109/ICMLA.2018.00235

Katsumi, M., &Fox, M. (2018). Ontologies for Transportation Research: A Survey. *Transportation Research Part C: Emerging Technologies*, *89*(September 2017), 53–82. https://doi.org/10.1016/j.trc.2018.01.023

Kim, D. S., &Yoon, W. C. (2013). An accident causation model for the railway industry: Application of the model to 80 rail accident investigation reports from the UK. *Safety Science*, *60*, 57–68. https://doi.org/10.1016/j.ssci.2013.06.010

Kim, S., &Yeo, H. (2017). Evaluating link criticality of road network based on the concept of macroscopic fundamental diagram. *Transportmetrica A: Transport Science*, *13*(2), 162–193. https://doi.org/10.1080/23249935.2016.1231231

Kulkarni, A., &Shivananda, A. (2019). Natural Language Processing Recipes. In *Natural Language Processing Recipes*. Apress. https://doi.org/10.1007/978-1-4842-4267-4

Li, J., Sun, A., Han, J., &Li, C. (2020). A Survey on Deep Learning for Named Entity Recognition. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*. http://neuroner.com/

Li, X., Shi, T., Li, P., Yang, L., &Ma, X. (2018). BiLSTM-CRF model for named entity recognition in railway accident and fault analysis report. *ACM International Conference Proceeding Series*, *Part F1482*, 1–5. https://doi.org/10.1145/3321619.3321623

Rausand, M. (2013). *Risk assessment: theory, methods, and applications* (John Wiley & Sons. (ed.); Vol. 115).

Young, T., Hazarika, D., Poria, S., &Cambria, E. (2017). *Recent Trends in Deep Learning Based Natural Language Processing*. http://arxiv.org/abs/1708.02709

Young, T., Hazarika, D., Poria, S., &Cambria, E. (2018). Recent trends in deep learning based natural language processing [Review Article]. *IEEE Computational Intelligence Magazine*, *13*(3), 55–75. https://doi.org/10.1109/MCI.2018.2840738

Yu, G., Zheng, W., Wang, L., &Zhang, Z. (2018). Identification of Significant Factors Contributing to Multi-attribute Railway Accidents Dataset (MARA-D) Using SOM Data Mining. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2018-Novem*, 170–175. https://doi.org/10.1109/ITSC.2018.8569336

Zhou, J. L., &Lei, Y. (2018). Paths between latent and active errors: Analysis of 407 railway accidents/incidents' causes in China. *Safety Science*, *110*(November 2017), 47–58. https://doi.org/10.1016/j.ssci.2017.12.027