Chapter 40
Big Data and Natural Language Processing for Analysing Railway Safety:
Analysis of Railway Incident Reports

Kanza Noor Syeda
Lancaster University, UK

Syed Noorulhassan Shirazi
Lancaster University, UK

Syed Asad Ali Naqvi
Lancaster University, UK

Howard J Parkinson
Digital Rail Limited, UK

Gary Bamford
Digital Rail Limited, UK

ABSTRACT

Due to modern powerful computing and the explosion in data availability and advanced analytics, there should be opportunities to use a Big Data approach to proactively identify high risk scenarios on the railway. In this chapter, we comprehend the need for developing machine intelligence to identify heightened risk on the railway. In doing so, we have explained a potential for a new data driven approach in the railway, we then focus the rest of the chapter on Natural Language Processing (NLP) and its potential for analysing accident data. We review and analyse investigation reports of railway accidents in the UK, published by the Rail Accident Investigation Branch (RAIB), aiming to reveal the presence of entities which are informative of causes and failures such as human, technical and external. We give an overview of a framework based on NLP and machine learning to analyse the raw text from RAIB reports which would assist the risk and incident analysis experts to study causal relationship between causes and failures towards the overall safety in the rail industry.

DOI: 10.4018/978-1-5225-8356-1.ch040

INTRODUCTION

In this chapter, we describe the research we have been undertaking to understand Big Data (BD) and its application to management of safety in the railway. We undertook the journey described in this chapter because we realised that many of the traditional safety management approaches do not deal very well with the complex socio-technological systems we are increasingly facing in the railway environment. We notice that we are in a new paradigm when it comes to BD, Internet of Things (IoT), computing power and intelligent algorithms. There is a vast potential to take a data driven approach to safety and systems engineering with decisions being based on real data and not just engineering judgments. Our aim is to help in the construction of a suite of BD risk assessment and development tools for reducing safety risk in railway projects and operations. In the first part of the chapter we present a summary of our previous works (Angelov, Manolopoulos, Iliadis et al., 2016; Parkinson & Bamford, 2016; Parkinson, Bamford, & Kandola, 2016; Parkinson & Bamford, 2017) in which we set out to develop an understanding of BD as it applies to railway safety management. It also helps set the scene for the NLP research which is the main focus of the rest of this chapter.

In order to understand what was meant by BD and its application in railway safety, we undertook an initial phase of research. This research involved an investigation into various railway accident to explore accident causation and assess if the available data could have provided a prior warning of the catastrophe. This research also included evaluating whether the assessment could be classified as a BD approach or simply business as usual (BAU). We proposed a new mechanism for identifying and mitigating heightened risk called ELBowTie[1] (Parkinson et al., 2016). The next stage of our research was to go into a deeper analysis of the Grayrigg Accident (Branch, 2011). We analysed the engineering and management failures associated with Grayrigg using a bowtie risk assessment approach. We investigated the type of Big Data Analytics (BDA), available, that could potentially have been used to identify hazardous conditions prior to the accident. We then undertook meta-analysis conducted in previous research in order to develop an understanding of the wider state of play of intelligent analytics in current railway research and development. We finally move on the main focus of the chapter which describes on-going work being undertaken to employ machine learning and Natural Language Processing (NLP) to predict railway heightened risk.

The field of incident analysis consists of number of methods. Certain methods are based on accumulated expert knowledge with prescribed models and/or procedures. Although, these methods differ amongst themselves in terms of their level of detail, methodology, presumptions, aspects of focus, etc.; most prescribe certain basic Entities of Interest (EOI) that maybe common within several methods. We define EOI as factors that represent categories of information that may help explain an incident in terms of cause-effect relationships. Furthermore, due to the heterogeneous nature of each incident, a lot of relevant information is recorded in loose text instead of constrained value fields. Such text components enclose considerable richness that is invaluable for incident analysis and prediction. However, there is only limited work available aimed at applying text analysis to incident investigation. This is, primarily due to the difficulty and challenges related to interpretation of such data.

Natural Language Processing (NLP) represents a set of techniques that can computationally extract useful conceptual information from text. Our goal in this study is to assess the usefulness of NLP to the field of incident analysis in terms of identifying EOI from incident analysis reports. More specifically, based on that understanding, we want to ascertain the usefulness of NLP approaches to examine the presence of significant entities which are based on expert knowledge and presence of relationships among