Exploiting Natural Language Processing for Analysing Railway Incident Reports

Kanza Noor Syeda[1], Syed Noorulhassan Shirazi[2], Syed Asad Ali Naqvi[2], and Howard J Parkinson[3]

[1] Department of Entrepreneurship, Strategy & Innovation, Lancaster University Management School, Lancaster University, LA1 4YX, UK
kanzanoor@gmail.com
[2] InfoLab21, School of Computing and Communications, Lancaster University, LA1 4WA, UK
{n.shirazi, s.naqvi}@lancaster.ac.uk
[3] Digital Rail, Knowledge Business Centre, Lancaster University, Lancaster, LA1 4WA, UK
hjparkinson@digitalrail.co.uk

Abstract. In this work, we focus accident causation for the railway industry by exploiting text analysis approaches mainly Natural Language Processing (NLP). We review and analyse investigation reports of railway accidents in the UK published by the Rail Accident Investigation Branch (RAIB), aiming to unleash the presence of entities which are informative of causes and failures such as human, technical and external. We give an overview of framework based on NLP and machine learning to analyse the raw text from RAIB reports which would assist risk and incident analysis experts to study causal relationship between causes and failures towards overall safety in rail industry. The approach can also be generalized to other safety critical domains such as aviation etc.

Keywords: Railway safety, rail, incident analysis, accident causation model

1 Introduction

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 each other 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 a number of methods. We define EOI as factors that represent categories of information that may help explain an incident in terms of cause effect relationships. Further due to the heterogeneous nature of each incidents, plenty of relevant information is recorded in loose text instead of constrained value fields. As a result, such text components enclose considerable richness that is invaluable for incident analysis and and prediction. However, there is scarcity of work that aim to apply text analysis for incident investigations 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 those entities. This work draws attention to the numerous opportunities available by exploiting these technologies to enhance railway safety.

1.1 NLP for Railway Incident Analysis

Incident analysis reports usually comprise of an assorted set of information from incident description to legal obligations, to policies, procedures, and description of various components. Much of the information in these reports represents tangential issues with little relevance to incident factors.

Further, information related to incident factors may be scattered across the documentation. In this context NLP approaches need to be honed towards the identification of relevant information – incident factors and their relationships. What is needed is a point of reference for the NLP approaches to conduct their search. This point of reference was provided through a set of general concepts that are relevant to incident analysis. We have termed these concepts as Entities of Interest (EOI). EOIs (see Table. 1) are abstract concepts that represent different types of faults and causal factors in incident analysis. In deriving EOIs our goal was to cover a broad spectrum to incident relevant concepts. The broader the spectrum of EOIs, the lesser the possibility that NLP analyses based on these EOIs will miss any relevant incident factors.

1.2 EOI Extraction

To derive the list of EOIs we turned to the literature on incident analysis [13, 14, 10, 8, 15, 24, 11, 21, 25]. Many mature incident analysis methods express reality by means of a limited number of prescriptive models, artefacts, and taxonomies. These prescriptive models and artefacts represent element of accumulated expert knowledge and experience. These methods explain incidents through a number of abstract incident types and causes, for example, the list of general failure types in Tripod, analysis tree in MORT, and SORTM in HPIP. We identified a number of significant incident analysis methods within this category and used their models, artefacts, and taxonomies to inform the derivation of our EOIs. Our criteria for selecting these incident analysis methods were a) they had to be generalised (not domain specific) in their application, and b) they had to be significant in the field. Significance means that the method should have been used in multiple studies by different authors, or that it is widely used in industry for incident analysis. The EOIs in relation to the incident analysis methods are shown in Table 1 and Appendix A.

Moreover, the work intends to inform and advance railway incident analysis efforts by illustrating how these novel technologies, together with expert knowledge, can be leveraged in order to improve railway safety. In particular, we are interested in extracting entities of interest from textual report describing incidents. In doing so, we hope to unleash causes of accidents, catalogue potential safety measures, and show how to better leverage technologies to improve the safety. The work mainly involves but not limited to the following:

- Review state-of-the-art incident analysis techniques with respect to railway safety
- Understanding various NLP techniques for named entity recognition
- Classification and prediction approaches based on machine learning to establish correlation between causes and failures.
- Design and build tool chain using Python and Perl scripts to provide implementation of our model

The structure of the remainder of this paper is: Section 2 details background information by providing a purpose and benefits of incident analysis. Section 3 discusses the method and details overview of the process. Section 4 presents our analysis and Section 5 concludes the paper.

2 Background

2.1 Incident Analysis

Accidents, and incidents of faults and failures are an unavoidable reality for even moderately complex systems. Accidents, though unfortunate events, also provide an opportunity to uncover vulnerabilities and latent errors in systems. In this vein accident and incident analysis plays an important role in improving system dependability and robustness. The purpose and benefit of an incident analysis is realised when it is able to explain the dynamics of "how" and "why" an incident happened in terms of its root causes and other relevant incident factors. Root causes are those entrenched systemic risk factors within an enterprise, environment, or domain that may have either caused or contributed to a number of latent and active failures. In [20] states that accidents happen when latent and active failures converge under the right circumstances. Therefore, unless the root causes are identified and their effects mitigated, the risk of accidents in a domain will persist as before. Conversely, the identification and mitigation of the root causes will contribute towards the avoidance of incidents in the future. We define incident factors as all relevant factors related to an incident, which may include causal factors, Performance shaping factors, Post-event factors, and other matters concerning safety.

2.2 Multi-incident Analysis

Incidents when analysed individually often seem to be caused due to isolated reasons such as equipment failure, mishap, or negligence. However, when incidents are analysed in the context of other incidents in the broader domain then patterns begin to emerge between them. These patterns may indicate basic and underlying reasons for incidents the root causes. The practice of analysing a number of incidents together is called Multi-incident analysis. It provides the analyst with a broader view of the hazard space in a domain, which leads to lessons-learned that contribute towards improving the state of the art for the whole domain.

For the purposes of learning from and preventing future safety incidents, many fields have developed domain specific multi-incident analysis approaches. For example, the DATIX incident reporting system is used by the NHS in the UK for incident reporting and learning [1]; the Australian Incident Monitoring System (AIMS) is an incident reporting and analysis system for the health care sector in Australia [2]; AERO, AQD, BASIS, HeliStat, AirFASE, PEAT, QUORUM Perilog, Aviation Safety Data Mining Workbench are all multi-incident analysis and reporting system developed in the aviation industry [7].

The rail industry in Great Britain suffers around 75,000 safety related mainline railway incidents. To learn from these incidents, they are recorded in a database called the Safety Management Information System (SMIS). These events that cover everything from derailments and signals passed at danger to passenger slips, trips and falls and operating irregularities, are classified through the Incident Factor Classification System (IFCS) [9].

Most multi-incident analysis methods, including the aforementioned, are quantitative in nature. Generally, these methods collect data about incidents into a database through an incident reporting system. Then use statistical analysis to find interesting correlations between various factors, fields, and characteristics of the data. However, there are a number of inherent problems in this type of analysis which may make it difficult to discover root causes, these include:

- **Inappropriate abstraction**: The attributes used to describe an incident would have to be reasonably generic so as to accommodate a large number of incidents. Consequently, the distinct details of each incident are abstracted away due to this generic representation.
- **Irrelevance**: Generally the choice of attributes is not specifically to support root cause analysis, but may be influenced by regulatory requirements, statistical reporting requirements, and also the convenience and objectivity with which an attribute may be measured.
- **Semantic differences**: The changing definitions of attributes across time and space and also due to the different purposes of the data and collection methods used, may create data integration and normalisation problems for statistical analysis.

Further, the data points or incident factors being recorded in the database are chosen by the creators of the database according to their accumulated insights or domain knowledge at the time. As the domain evolves some new incident factors may become relevant, and some old ones may become irrelevant. Therefore, there is a need to periodically update the domain knowledge and the incident factors, so that the multi-incident database remains relevant and useful. Domain knowledge about the dynamics of incidents and accidents is implicit within incident analyses reports. These reports can, therefore, be used as source material for updating the multi-incident database.

In our work we have used various Natural Language Processing (NLP) techniques to jointly analyse incident analyses reports from the domain of Railway Incident Analysis, to identify interesting patterns of incident factors and the relationships between them.

Table 1: Comparison of incident analysis

EOI	**Description**	**Tripod**	**FishBone**	**TOR**	**HPIP**	**CREAM**	**MORT**
EF	Equipment Failure (Hardware)	Hardware	Machines	Property loss (breakage or damage)	–	Equipment Failure	–
MP	Maintenance problem	Maintenance management	Measurements	Supervision (unsafe acts, initiative)	–	–	Maintenance
DP	Design problem	Design	Material	–	–	–	Design
OP	Operational Procedures	Operating procedures	Methods	Operational (job procedures)	Procedures	–	Procedures
VC	Violation Conditions	Violation including conditions	Environment	–	–	–	–
–	–	House Keeping	–	–	Human Engineering (Bad lights, errors not detectable)	–	–
–	Incompatible goals	–	–	–	–	–	–
CF	Communication factors	–	Authority (bypassing, conflicting orders)	Communication	–	Communication	
OF	Organizational factors	Organization	–	Management (policy, goals)	Organization factors	Management problems	Management
TF	Training factors	Organization	–	Coaching (unusual situation, training)	Training	Lack of knowledge	Training
–	Defence planning	–	–	–	Inadequate plan	–	
HF	Human Factor	–	People	Personal traits	–	–	Human factors

3 Method

To this end we focus on the domain of Railway Incident Analysis, and propose a framework that can bridge the gap between text analysis and incident analysis. The following are key tasks (see Figure. 1 for more details):

- Obtain a number of railway incident analysis reports as raw corpus - this is done through a custom crawler scripts that extracts summary of 298 reports from RAIB[4].
- The corpus is divided into sets of training and testing textual data for subsequent phases of the process.
- Produce a list of appropriate EOI from existing incident analysis approaches literature, such as Tripod, MORT etc (see Table. 1).
- The EOI list will be used to manually extract relevant text representing each EOI from a number of incident analysis reports from the railway incident domain.
- The text thus extracted are processed using textual analysis and NLP and transformed into representative vectors.
- These feature vectors are fed as input to a machine learning based classification system, for training purpose.
- The testing textual data is also processed using NLP and supplied to the trained classification system that identifies EOI within the testing data.
- The output of the classification system may be used in subsequent analyses such as visualisation, clustering , co-occurrence and correspondence etc.

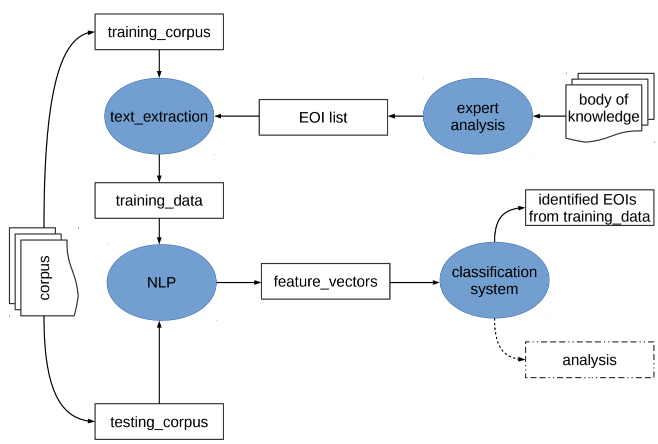


Fig. 1: Process overview.

[4] https://www.gov.uk/raib-reports?report_type\%5B\%5D=investigation-report

3.1 Data

The investigation reports of railway accidents in the UK published by the Rail Accident Investigation Branch (RAIB), the independent railway accident investigation organization for the UK, from 2008 to 2010. These reports are quite comprehensive and provides information related to an incident in detail, such as sequence of events preceding, during, and following the accident; the immediate cause, causal factors, and other factors contributed to the accident. Furthermore, it provides recommendations to address the causal and contributory factors.

4 Analysis

Traditional text analysis approaches are based on using a structure and language consistent with first-order predicate logic where data is represented in terms of tuples and grouped into relations [4]. A problem arises when such approaches are used to analyse the incident analysis reports because mostly these reports include textual data that is impossible to analyse with the traditional approaches such as relational databases. In particular, when it comes to identify the patterns and define sub categories that would enable incident analysts and experts to categories the entities meaningfully. Below we discuss text analysis approaches which potentially can be useful for domain of railway incident analysis.

4.1 Topic Modeling

Topic modeling is a text analysis technique which enable to discover the main themes that pervade a large and otherwise unstructured collection of textual data. It can organize the collection according to the discovered themes and can be adapted to many kinds of data - such as extracting information from incident reports [3]. This technique has previously been applied successfully in order to estimate the expected duration of the incidents i.e, defined as tie period that spans fro incident occurrence to clearance [18, 17, 16]. This period allow the experts to plan and execute a response strategy. We believe this can still be very much relevant to railway incident analysis. However, our aim in this work is to analyse the text from existing reports in order to reveal the key entities of interest and possibly understand the relationship between those entities through quantified means. Such quantified relationship between entities can provide guidance to incident experts when they perform post event analysis.

Here, we analyse subset of our main corpus consist of six incident reports and used a topic modeling technique to infer the hidden topic structure. We then computed the inferred topic distribution (see Table. 3), the distribution over topics that best describes its particular collection of words. Our results show the most probable entities/topics which can be useful to build the causation model by mapping each entity against cause and effect category. Therefore, these interpretable entities arise by computing the hidden structure that likely generated the observed collection of incident reports. For example, Table. 2 lists topics discovered from underlying reports (i.e, subset of corpus).

The usefulness of topic models for railway incident analysis is due to the fact that the inferred hidden structure resembles the thematic structure of the collection which annotates each document in the collection. This could be taxing to perform manually during incident analysis process. Further, these annotation can be used to aid subsequent stages of incident analysis process like like information retrieval, classification and corpus exploration [3, 22].

Table 2: List of top 5 entities per document

Report ID	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	train	rail	track	signal	crossing
2	driver	inspection	buckle	points	trail
3	accident	bar	reported	signaler	around
4	switch	stretcher	temperature	mom	level
5	brake	work	day	box	tractor
6	door	permanent	derailed	danger	approaching

For this particular task, we use `topicmodels`[5] package in `R`[6] which provides implementation of LDA[7] technique. LDA facilitates the automatic identification of entities from corpus. The aim is to infer the latent structure of entities given the words and documents. The way LDA works is that for each topic (number of topics are predefined - 5 in our study) two actions are performed. First, it computes the proportion of words in document d that are currently assigned to topic t, and secondly, the proportion of assignments to topic t over all documents that come from this word w. It then reassign w a new topic based on generative model which is essentially the probability that topic t generated word w, so it makes sense that we re-sample the current word's topic with this probability. The process is repeated until it reaches a steady state where all assignments are good.

However, an important step in text analysis is preprocessing and we perform this preliminary step in order to transform the corpus for analysis. The `tm` package is used to perform clean up (removing hyphens and colons etc), remove punctuation, convert to lower case, remove common words (such as articles, conjunction and common verbs etc), stemming[8] and lemmatisation[9]. The later is needed to takes grammatical context into account. The Table 2 and Table 3 below list the top terms in topics 1 through 5 and topic probabilities respectively. The highest probabilities are shown in bold.

[5] https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf
[6] https://www.r-project.org/about.html
[7] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
[8] https://en.wikipedia.org/wiki/Stemming
[9] https://en.wikipedia.org/wiki/Lemmatisation

Table 3: Topic probabilities

Report ID	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	**0.2039801**	0.07462687	0.08955224	0.09950249	**0.53233831**
2	**0.28823529**	0.11176471	0.10588235	0.13529412	**0.35882353**
3	**0.2056962**	**0.48417722**	0.11708861	0.09493671	0.09810127
4	**0.41441441**	0.12612613	0.16216216	0.11711712	0.18018018
5	0.13846154	0.13846154	**0.49230769**	0.07179487	0.15897436
6	0.11911357	0.09695291	0.09418283	**0.65096953**	0.03878116

The quick analysis of the two tables shows that the technique has performed well. For instance, topic 5 is about level crossing that caused that accident. Similarly, the report 6 is about derailment. The highest probability in each row is in bold. By looking at these results, it seems that incident expert can use such a technique to perform unsupervised classification of corpus of documents which can assist in subsequent incident analysis process. However, it is important to examine the results carefully to check the validity before inferring analysis.

4.2 Cluster Analysis

Generally, the problem incident analyst face is how to catgorise large collection of documents in some meaningful way. The problem is due to the fact that such reports generally do not have predefined classification schemes that is known to fit the collection. Therefore, techniques like clustering analysis can also be useful to analyse these reports automatically based on their structure and content. It is worth mentioning here that clustering results depend rather critically on the underlying algorithm that is employed [5]. For cluster analysis we use main dataset (`corpus294` - i.e. summary of 298 RAIB reports) and subset of it with 6 reports (`corpus6` - containing text against entities of interest) and load into an object that can be manipulated by `tm` package. After the preprocessing steps as explained in previous §Section 4.1 we create a Document Term Matrix (DTM)- i.e., a matrix in which the documents are represented as rows and words as columns. For `corpus6` there are six documents and nearly 2000 words which can be mathematically represented as 2000 dimensional space in which each of the word representing a coordinate axis and each document is represented as a point that space (illustration is shown in Figure. 2).

We employ hierarchical clustering using Ward's method [23] in order to compute distance between these documents. The visualization of these grouping for both corpora is shown in Figure. 3. Each branch represents a distance at which a cluster merge occurred.

Clearly, the close branches means high similarity. From incident analysis point of view we can obtain how closely two reports similar thus potentially be indication of similar incidents. This could also mean the flow of events between those two incidents is highly correlated. Such an information can be useful for

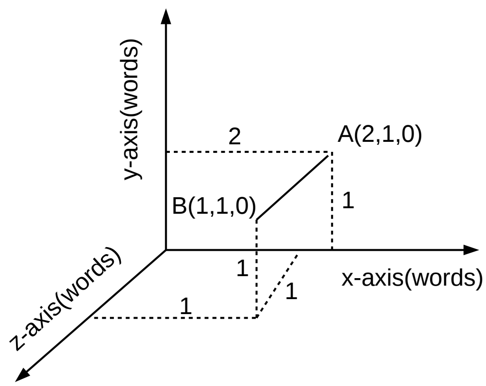


Fig. 2: Illustration of document vectorization in 3-word space

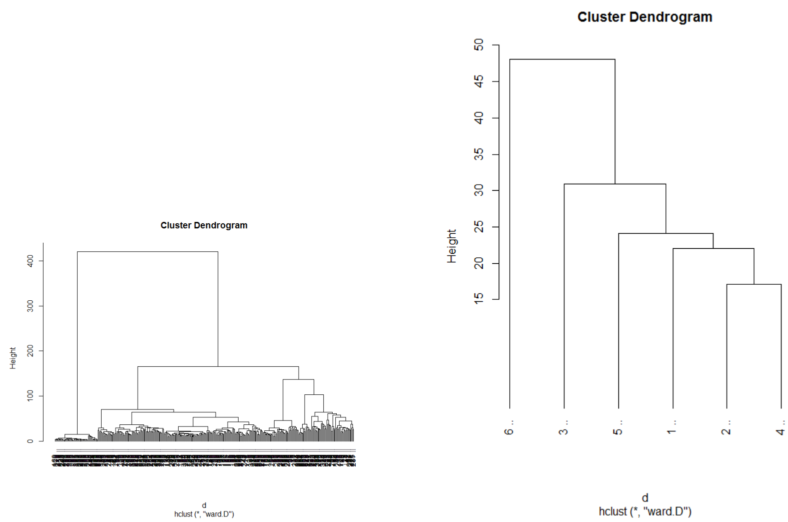
incident experts when complementing evidence is obtained by quantifying similarity of incident reports. Similarly, we plot the clusters in Figure. 4 for both corpora and, the intuition behind these plot is to apply dimensionality reduction technique for ease of visualization.

The dimensionality reductions capture the variability between the cluster and plot first two components - which in our case explains 70% variability for `corpus298` and over 13% for `corpus6`. However, the results do not have straightforward interpretation when it comes to analysing incident reports. However, there are many text mining techniques that perform better in grouping based on entities rather than word frequencies.

4.3 Natural Language Processing

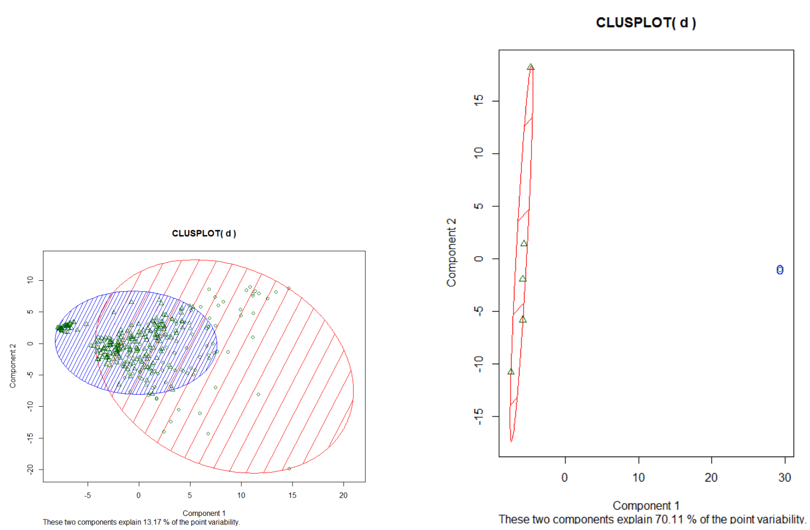
In order to support the incident analysis, we investigated the natural language processing techniques to get an insight by converting the text into spatial representation of network of words and entities. These graphs of entities visually express the relationship between words and entities of interest which can provide an understanding how the entities influences safety on railways. In this work we were focused on visual representation of the key entities and their relationships. Mainly there are three main approaches when it comes to extraction of information from text: thematic, networks and semantic [19]. The topic modeling approach presented earlier is kind of thematic approach since it is based on the frequency of concepts that allows classification of the topics of text. On the other hand semantic analysis also takes into account the relationships among concepts using semantic grammar, while network analysis is based on network text analysis to obtain semantically linked concepts. Similar to this approach the authors in [6] analyse Close Call Records[10] to get insight into railways safety. In order to test value of NLP we preprocess the dataset, `corpus6` through *tagging*

[10] `https://www.rssb.co.uk/risk-analysis-and-safety-reporting/reporting-systems/close-call-system`



(a) Dendrogram for `corpus298` (b) Dendrogram for `corpus6`

Fig. 3: Cluster grouping



(a) Cluster grouping for `corpus298` (b) Cluster grouping for `corpus6`

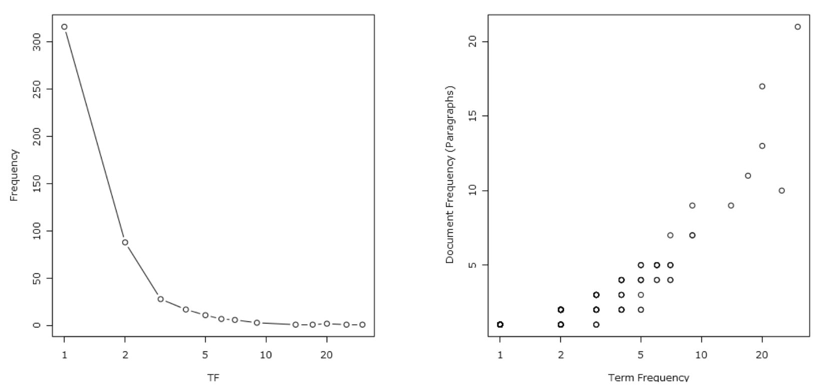
Fig. 4: Variation explained by principal components

and *tokenization* process as explained in [12]. We perform these analysis using

KHCoder[11] which is a free software for quantitative analysis and computational linguistics.

The main objective of using NLP approach for incident analysis in this work is to matching of *entireties of interests* which are specified using rich linguistics patterns that incorporates as fundamental the notion of embedding of patterns and various linguistic predicates. Further, this would help to analyse the structure of words, phrases and sentences (making use of linguistic rules).

To meet that objective, We segment the extracted text file related to each EOI into words using, *stemming*, *lematization* and *Part of Speech (POS)* tagging. The later divides the data into the simples POS such as verbs, and nouns and includes conjugated forms ad distinct entities. The descriptive statistics including term frequencies and document frequency distribution are plotted. These statistics visually represents the number of documents that contains each term and also to evaluate the correlation between term frequency, the number of occurrences of each term in the data, and document frequency, the number of documents in which each term is used (see Figure.5).



(a) Term frequency distribution

(b) Term frequency-document frequency plot

Fig. 5: Descriptive statistics of words

In order to find the words that closely associated with the *entities of interests*, we perform word association by making estimations from data using conditional probability calculations. For example which words closely related to the entity like *Human Factor* or Technical Factor. The Figure 6 shows a correspondence analysis on extracted words to explore what kinds of word have a similar appearance pattern. The top right shows the words related to EOI of *Hardware/Equipment failures*. Such an information can help analyst to build the causation model by identifying causes and failures through related terms.

[11] `http://khc.sourceforge.net/en/`

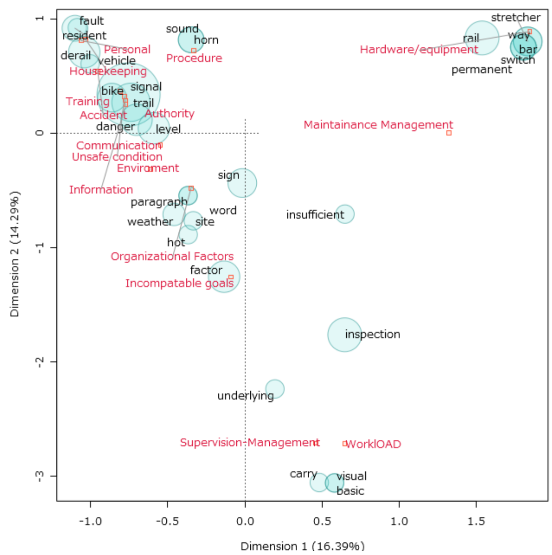


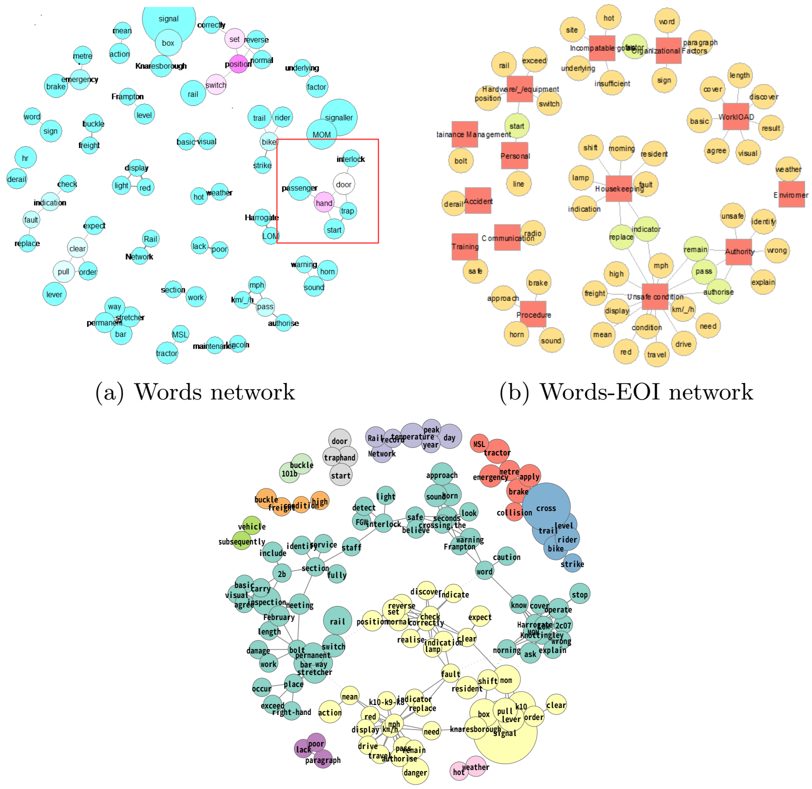
Fig. 6: Results of a correspondence analysis

Similarly, the network diagram is shown in Figure. 7 which shows the words with similar appearance patterns i.e., with high degree of co-occurrence, connected by edges. This helps to understand the co-occurrence structure of words. In addition, the entities to text mapping is also shown which represents the association between entities and words. For instance, the sequence of events such as "*a passenger hand being trapped in the door interlocking mechanisms*", highlighted in red. The correspondence and association analysis is solely drawn from a text file which contains the representative text of against each EOI from six incident reports.

5 Conclusions

This paper describes the systems that combines Natural Language Processing (NLP) to obtain the capability to gain insight into railway accident reports by unleashing previously unknown patterns of interests for railway safety. We show that it is possible to gain high value by using text analysis to map a different sequences of language text to the concepts and entities in incident domains. The frequency, distribution and co-occurrence of these entities form patterns can provide useful indicators for investigations, and assist incident experts in establishing root cause analysis using relevant supporting information.

Further, such a process can be enriched as when more data is available by applying additional machine learning approaches. The enriched data such as investigation reports not only contains part of speech information, but also a rich lexicon, and a great deal of domain knowledge embodied on concepts and



(a) Words network (b) Words-EOI network

(c) Modularity co-occurrence

Fig. 7: Co-occurrence network

entities. This knowledge can be fed into machine learning to effectively provide means to express concepts and entities for advances of safety in railways.

Availability This raw dataset and our custom scripts on which this work is based, is available on line[12].

Acknowledgments This work is sponsored by Santander Universities program and Digital Rail. We are also thankful to RAIB for providing the railway accident investigation reports on which this work is based. We are also grateful to RSSB for providing insightful comments and input to the use of incident classification system. We are also thankful to Paul Rayson and Stephen Wattam for useful discussion on Natural Language Processing approaches.

[12] https://github.com/kanzanoor/nlp-paper-digitalrail.git

Appendix A Entities of Interest (EOI)

Table 4: Description of EOI for each method

Model	EOI	Abbreviation	Description
Tripod	Hardware	HW	Inadequate supply and function of equipment and materials.
Tripod	Maintenance management	MM	A necessary maintenance activity was delayed or postponed.
Tripod	Design	DE	An inherent design problem in a device that may cause the device to function in an unexpected manner
Tripod	Procedures	PR	Procedures may be ambiguous, incorrect, or inappropriate
Tripod	Violation-Inducing conditions	VC	Conditions that pressure the workers to violate procedures.
Tripod	Housekeeping	HK	Conditions and violations that are known, but have not been fixed over a period of time.
Tripod	Incompatible goals	IG	Organisational and individual goals that conflict with safety requirements.
Tripod	Communication	CO	A message could not be sent, sent to the wrong recipient, or misinterpreted by the recipient.
Tripod	Organization	OR	Organisational structures may hinder the prevention or mitigation of incidents
Tripod	Training	TR	Lack of competent workers due to inadequate training.
Tripod	Defence planning	DP	Deficiencies in the planning of detection and response procedures for incidents.
FIshBone	People	PE	Anyone involved in the incident
FIshBone	Methods	MD	The processes including policies, procedures, rules, regulations and laws
FIshBone	Machines	MC	Any equipment, computers, tools, etc.
FIshBone	Materials	MT	Raw materials, parts, pens, paper, etc.
FIshBone	Measurements	MS	Data used in the process
FIshBone	Environment	EN	Conditions, such as location, time, temperature, and culture
TOR	Unusual situation	US	failure to coach(New employee,tool,equipment,process,material)
TOR	Training	TR	Not formulated or need not foreseen
TOR	Conflicting Goals	CG	duties and goals are not clear
TOR	Authority	AU	Authority inadequate to cope the situation
TOR	Co-operation	CO	failure to plan(co-operation)
TOR	Supervision	SP	failure to see problems,observe and correct
TOR	Property loss	PL	Accidental breakage or damage
TOR	Unsafe conditions	UC	Inefficient or unsafe conditions
TOR	Personal traits	PT	Physical condition,Impairment,Alcohol,personality,work Habits and work assignment
HPIP	Procedures	PR	Formal written guidance provided to workers or supervisors.
HPIP	Training	TR	Training indicated as if an individual failed to perform
HPIP	Organizational Factors	OF	Factors that influence human performance reliability and enhance organizational effectiveness
HPIP	Work environment	WE	Housekeeping,bad lights,cold/hot,noisy
HPIP	Human Engineering	HE	Reliable human performance(Human-machine interface,complex systems, Non-Fault Tolerance System
HPIP	Supervision	SP	Inadequacies in task planning and follow-up contributed to an event
HPIP	Communication	CO	Misunderstood,late communication,no communication

Table 5: Description of EOI for each method

Model	EOI	Abbreviation	Description
IFCS	Practice and processes	PP	The rules, standards, processes and methods of working
IFCS	Communication	CO	How we relay information to each other in the context of safety critical information.
IFCS	Information	IN	Information is used to support an activity
IFCS	Workload	WL	Workload is about understanding the demand created by particular activities
IFCS	Equipment	EQ	Faulty,design not compatible with its use
IFCS	Knowledge skills experience	KE	appropriate knowledge,familiar with the circumstances.
IFCS	Supervision-management	SM	Decisions about resources, budgets, work allocation and planning
IFCS	Work environment	WE	lighting levels, noise, temperature and vibrations
IFCS	Personal	PE	a collection of influences that may affect the individual(fatigue,physical and mental well-being)
IFCS	Team Work	TW	How to work together and coordinate to achieve safe performance
CREAM	Working conditions	WC	
CREAM	Operational support	OS	support provided by specifically designed decision aids
CREAM	Procedures	PR	formally define patterns of response, heuristics, or routines to be used
CREAM	Available time	AT	time to deal with the situation
CREAM	Training and experience	TE	operational experience, training, or familiarization
CREAM	Collaboration quality	CQ	social climate among the workers
MORT	Communication	CO	
MORT	Maintenance	MM	
MORT	Design	DE	
MORT	Operability	OP	
MORT	Training	TR	
MORT	Procedure	PR	
MORT	Time	TI	
MORT	Knowledge	KN	
MORT	Worker Problem	WP	
MORT	Technical Information System	TI	

References

1. Alrwisan, A., Ross, J., Williams, D.: Medication incidents reported to an online incident reporting system. European journal of clinical pharmacology 67(5), 527–532 (2011)
2. Beckmann, U., Baldwin, I., Hart, G., Runciman, W.: The australian incident monitoring study in intensive care: Aims-icu. an analysis of the first year of reporting. (1996)
3. Blei, D.M.: Probabilistic topic models. Communications of the ACM 55(4), 77–84 (2012)
4. Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM 13(6), 377–387 (1970)
5. Diday, E., Simon, J.: Clustering analysis. In: Digital pattern recognition, pp. 47–94. Springer (1980)
6. Figueres-Esteban, M., Hughes, P., Van Gulijk, C.: The role of data visualization in railway big data risk analysis. In: European Safety and Reliability Conference (2015)
7. GAIN Working Group B, A.M., Tools: Guide to Methods and Tools for Airline Flight Safety Analysis (2003)
8. Gano, D.L.: Reality Charting: Seven Steps to Effective Problem-solving and Strategies for Personal Success. Apollonian Publications (2011)
9. Gibson, H., Smith, S., Lowe, E., Mills, A., Morse, G., Carpenter, S.: Incident factor classification system. Rail Human Factors Supporting Reliability, Safety and Cost Reduction. Taylor & Francis Group, London, UK pp. 683–691 (2013)
10. Harms-Ringdahl, L.: Guide to safety analysis for accident prevention. IRS Riskhantering (2013)
11. Hollnagel, E., Speziali, J.: Study on developments in accident investigation methods: a survey of the state-of-the-art. ski report 2008: 50, ski (2008)
12. Hughes, P., Figueres-Esteban, M., Van Gulijk, C.: Learning from text-based close call data. In: Safety and Reliability. pp. 1–15. Taylor & Francis (2016)
13. Johnson, C.: A handbook of incident and accident reporting (2003)
14. Latino, R.J., Latino, K.C., Latino, M.A.: Root cause analysis: improving performance for bottom-line results. CRC press (2016)
15. Livingston, A., Jackson, G., Priestley, K.: Root causes analysis: Literature review. HSE Contract Research Report (2001)
16. Noland, R.B., Polak, J.W.: Travel time variability: a review of theoretical and empirical issues. Transport reviews 22(1), 39–54 (2002)
17. Ozbay, K.M., Xiao, W., Jaiswal, G., Bartin, B., Kachroo, P., Baykal-Gursoy, M.: Evaluation of incident management strategies and technologies using an integrated traffic/incident management simulation. World Review of Intermodal Transportation Research 2(2-3), 155–186 (2009)
18. Pereira, F.C., Rodrigues, F., Ben-Akiva, M.: Text analysis in incident duration prediction. Transportation Research Part C: Emerging Technologies 37, 177–192 (2013)
19. Popping, R.: Computer-assisted text analysis. Sage (2000)
20. Reason, J.: Human error. Cambridge university press (1990)
21. Sklet, S.: Methods for accident investigation. Trondheim: Gnist Tapir (2002)
22. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning. pp. 977–984. ACM (2006)

23. Ward Jr, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association 58(301), 236–244 (1963)
24. Woloshynowych, M., Rogers, S., Taylor-Adams, S., Vincent, C.: The investigation and analysis of critical incidents and adverse events in healthcare (2005)
25. Ziedelis, S., Noel, M.: Comparative analysis of nuclear event investigation methods, tools and techniques. European Commission, Joint Research Center (EC/JRC), EUR 24757 (2011)