Text Mining-Based Association Rule Mining for Incident Analysis: A Case Study of a Steel Plant in India

Sobhan Sarkar\textsuperscript{1,5(✉)}, Sammangi Vinay\textsuperscript{2}, Chawki Djeddi\textsuperscript{3,4}, and J. Maiti\textsuperscript{5}

1 Business School, University of Edinburgh, 29 Buccleuch Pl, Edinburgh EH8 9JS, UK sobhan.sarkar@ed.ac.uk
2 Department of Mechanical Engineering, IIT Kharagpur, Kharagpur 721302, India
3 Department of Mathematics and Computer Science, Larbi Tebessi University, Tebessa, Algeria
4 LITIS Lab, University of Rouen, Rouen, France
5 Department of Industrial and Systems Engineering, IIT Kharagpur, Kharagpur 721302, India

Abstract. Although a large amount of accident data in terms of categorical attributes and free texts are available across large enterprises involving high-risk operations, the methodology for analyzing such mixed data is still under development. The present study proposed a new methodological approach to extract useful inherent patterns or rules for accident causation using association rule mining (ARM) of both incident narratives (unstructured texts) and categorical data. Incidents data from an integrated steel plant for a period of four years (2010–2013) are used for model building and analysis. In the first phase, the text mining approach is employed to find out the basic events that could lead to the occurrences of faults or incident events. In the second phase, text-based ARM has been used to extract the useful rules from unstructured texts as well as structured categorical attributes. A total of 23 best item-set rules are extracted. The findings help the management of the plant to augment the cause and effect analysis of accident occurrences as well as quantifying the effects of the causes, which can also be automated to minimize the human involvement.

Keywords: Text mining · Association rule mining · Occupational incidents · Steel plant

1 Introduction

Steel manufacturing sector is one of the high-risk sectors as the workers are exposed to hot, noisy, gaseous and/or dusty environment in most of the working hours [15]. From the study of [21], it is seen that metal workers take long-term sick leave which is more than double than that a normal worker takes.
Accidents (or undesired incidents) in such industry impose substantial burden on the stakeholders, both employers, and employees. According to the report of [11], the number of occupational accidents raise to 3 lakh mortalities, and 300 million injuries globally in every year. The cost of job related injury is more than $27 billion annually as per the estimation by the National Safety Council in USA. Although accident data are collected, they often remain either unutilized or under-utilized. Therefore, proper analysis of accident data is essential for identification of causal factors or faults and predicting their likely occurrences.

To understand the incident causation, incident data should be utilized properly. It is found from previous studies that most of the leading organizations try to maintain their own system of investigation and reporting of accidents. It eventually helps them take precautionary measures in order to lessen its re-occurrences. However, it is very difficult to use full information within incident investigation database due to the fact that they are mix of structured and unstructured, large and manually intractable incident data [3,28,40]. Incident data comprises different data types, such as categorical, continuous, and unstructured texts of incidents. The narrative field is aimed to provide a free text description of the incidents to elaborate its causes and contributory factors that otherwise remain hidden. In most of the published literature, it is observed that association rule mining (ARM) has been used on categorical incidents data; however, very few studies have been carried out which uses incident narratives [6]. Moreover, the texts data are so unstructured that creates a high degree of difficulty in pattern extraction from them. Therefore, there exists a need of research to extract incident patterns from structured and unstructured data and analyse them properly for accident analyses and prevention.

Therefore, the aim of the study is to extract meaningful patterns from both structured categorical and unstructured texts data using association rule mining (ARM) technique. Using this technique, rules or association of factors responsible for accidents are extracted. Moreover, predictive regions is used in text mining-based ARM which helps in generation of effective rules. Results from the study reveal that the proposed approach not only helps in automation of the process, but also provides the hidden rules that might be missed by expert’s knowledge. Therefore, preventive actions could be initiated in order to reduce the re-occurrence of incidents.

The rest of the paper is structured as follows: In the Sect. 2, related works are presented briefly. Dataset, data types, and the methods text-based ARM are described in Sect. 3. In Sect. 4, results are discussed. Finally, conclusions with limitations and scopes for future works are presented in Sect. 5.

2 Literature Review

Occupational incidents are serious concern for all industries. To minimize the number of incidents, investigation of causal factors is necessary. Some of the related studies also include prediction-based incident analysis and prevention, for examples, development of predictive model [34], rough set theory (RST) [29],
Bayesian Network (BN) [32], Petri nets [37], hybrid clustering [31], image processing [25], decision support system [30], optimized decision tree approach [35]. Most of these studies deploy data mining or machine learning-based techniques unlike traditional questionnaire-based analysis [33] due to their advantages over traditional ones. In addition, previous studies also indicated that the use of both categorical and text information eventually makes the analysis very difficult. Text mining (TM), in such cases, has been found to be an effective technique for analyzing such incident texts [6,26,27,39]. TM is usually used to find out the hidden patterns in the unstructured text fields which may be useful for improved information retrieval system or may be used as inputs to predictive models [6]. Overall, using TM in occupational accident analysis focuses on two broad domains: (i) auto-coding of texts [7,14,23,39], and (ii) text classification and prediction [6,9,26]. Auto-coding of incident reports reduce the human efforts to a great extent [4,14,39]. For example, [4] auto-coded compensation claims into two classes, i.e., musculoskeletal disorder (MSD), and slip-trip-fall (STF) using Naïve Bayes auto-coding classifier with 90% classification accuracy. Another work by [39] shows that Bayesian auto-coding model based on fuzzy and Naïve approaches could code near miss and injuries with considerable accuracy.

An effort is still necessary for information extraction for both unstructured text and non-text categorical data. ARM is a popular approach in this task, which has been applied in several safety studies, including construction [2,8,16], railway [18], steel [41], and roads [13,22]. The principal reason of using ARM is its capability of extracting useful patterns from huge datasets and its easier interpretation [2,17,20]. However, from the previous literature, it is found that most of the studies are conducted in construction sector. For instance, [16] tried to find associations of different factors as well as patterns of injuries in construction industry in Taiwan. [17] focused only on extraction of unapparent association rules using ‘extracted probability’. Although ARM has been used in those studies on categorical data, its application on both unstructured accident texts and categorical data is rather challenging and new. The text-based ARM approach has more potential to explore the hidden patterns from both types of data in terms of associations of different factors which most of the time remain under-utilized.

2.1 Research Issues

From the review of literature, the following research issues are identified:

(i) Large amount of text data or narratives of incidents remain often under-utilized across industries as it demands a substantial amount of human efforts and time, and often manually intractable.

(ii) Extraction of rules from incident narratives has not been addressed by any previous literature. Therefore, the use of texts is expected to have potential for providing hidden incident patterns or rules.

(iii) Finally, it is mostly observed from accident studies that ARM has been deployed in construction domains. Its use in other domains still remains unexplored which demands a thorough investigation.
2.2 Contributions of the Work

The study contributes to the body of literature as follows:

(i) Both the categorical and free-text data (i.e., narratives) have been used for generation of rules.
(ii) The present research encompasses both text mining-based rule-cause extraction using ARM. Predictive regions is used on text-based ARM for better rule extraction.
(iii) The proposed methodology has been validated using incident data obtained from an integrated steel plant in India.

3 Materials and Methods

In this section, a brief overview of the data set (i.e., variables and data types) obtained from an integrated steel plant and the methods used in this study are given below.

3.1 Variables and Data Types

The incident dataset for a period of 45 months (from 2010 to 2013) retrieved from the digital database system of an integrated steel plant was considered for the study. The dataset contains thirteen attributes (12 categorical, and one textual), and 998 observations. A brief description on each of the attributes is given below.

(i) **Date of incident (DOI):** This attribute implies the date when the incident was occurred.
(ii) **Department (Dept.):** This represents the location where the incident was taken place. In total, nine departments were considered, namely N1, N2, N3, N4, N5, N6, N7, N8, and N9. Here, N9 represents ‘other departments’ consisting of the departments having very few records of incidents during the study period.
(iii) **Incident outcome (IO):** It is the outcome variable. It has three different classes: (i) injury (I)-when someone gets injured physically from an incident; (ii) near miss (N)-when someone is narrowly escaped from an incident having full potential to cause injury or damage; and (iii) property damage (PD)-when there is damage to private or public property, due to the incident.
(iv) **Incident event (IE):** This attribute refers to the top event that qualifies the incident which has occurred. It has 23 classes. They are ‘crane dashing (CD)’, ‘dashing/collision (DC)’, ‘derailment (D)’, ‘electric flash (EF)’, ‘energy isolation (EI)’, ‘equipment/machinery (EM)’, ‘fire/explosion (FE)’, ‘gas leakage (GL)’, ‘hot metals (HM)’, ‘hydraulic/pneumatic (HP)’, ‘lifting tools & tackles (LTT)’, ‘process incidents (PI)’, ‘rail (R)’, ‘road incidents (RI)’, ‘run over (RO)’, ‘skidding (S)’, ‘slip/trip/fall (STF)’, ‘structural integrity (SI)’, ‘toxic chemicals (TC)’, and ‘working at heights (WAH)’.
(v) **Working Condition (WC):** This attribute represents the condition of work when the incident took place. It has three categories i.e., ‘group working (W1)’ representing the condition where people work in groups, ‘single working (W2)’ representing person working alone, and ‘Others (W3)’ representing situations when no workers were present.

(vi) **Machine Condition (MC):** It implies the condition of machine when the accident took place; either machine is in ‘idle condition (M1)’ or in ‘running condition (M2)’, or ‘others (M3)’ i.e., not related to machine.

(vii) **Observation type (OT):** This attribute represents the basic causes of incident and has four categories as; (i) ‘unsafe act (OT1)’ representing the person himself is responsible for the cause of incident, (ii) ‘unsafe act and unsafe condition (OT2)’ representing the incident occurred due to presence of both the factors, person’s fault and hazardous condition, (iii) ‘unsafe act by other (OT3)’ representing the incident occurred due to the other’s fault, and (iv) ‘unsafe condition (OT4)’ representing a situation which is likely to cause incidents.

(viii) **Incident type (IT):** This attribute represents whether an accident happened is due to ‘human behavior (IT1)’, or ‘process type (IT2)’ which is non-human fault.

(ix) **Standard Operating Procedure (SOP) Adequacy (SOPA):** SOP implies a procedure/guideline to be followed while performing tasks by the workers/operators. It has three categories: (i) ‘SOP adequate (SOPA1)’ – sufficient in quality and quantity; (ii) ‘SOP inadequate (SOPA2)’ – not sufficient in quality and quantity; and (iii) ‘not applicable (SOPA3)’.

(x) **SOP compliance (SOPC):** This attribute indicates whether any SOP was ‘followed (SOPC1)’, or ‘not followed (SOPC2)’, or ‘not applicable (SOPC3)’.

(xi) **SOP Availability (SOPAv):** This attribute implies that whether any SOP was ‘available (SOPAv1)’, or ‘not available (SOPAv2)’, or ‘not applicable (SOPAv3)’.

(xii) **SOP Requirement (SOPR):** This attribute represents that whether any SOP was ‘required (SOPR1)’, or ‘not (SOPR2)’, or ‘not applicable (SOPR3)’.

(xiii) **Brief description of incident (BDI):** This attribute consists of short description on how and why the incident occurred. The field contains free text logged by safety personnel after the incident was investigated.

### 3.2 Methods

The flowchart of the proposed methodology is depicted in Fig. 1. After data collection, text and non-text attributes are separated. Thereafter, text preprocessing techniques, namely tokenization, stemming & lemmatization, and stop words removal are performed. Finally, predictive regions are identified which finally produces pre-processed texts. Finally, rules are extracted from both pre-processed categorical and text attributes.
In order to find out the hidden pattern within data in terms of rules, ARM is used on both text and categorical data. The proposed text-based ARM can automatically discover association rules from text documents. The approach typically consists of three steps, which are explained as follows: (i) tokenization, filtration, stemming and indexing of the documents, (ii) determination of predictive regions in texts, and (iii) generation of association rules using ARM. All the three above-mentioned steps are briefly described below.

**Step 1 - Preprocessing of text:** Initially, the TDM has been obtained by text mining process. It is then indexed by proper weighting scheme i.e., Term Frequency-Inverse Document Frequency (TF-IDF), which is widely used in different literature [5,10,42]. The term TF-IDF can be expressed in the following Eq. (1):

$$w(i, j) = tf \times idf(d_i, t_j) = \begin{cases} N_{d_i, t_j} \times \log_2 \frac{|C|}{N_{t_j}} & \text{if } N_{d_i, t_j} \geq 1 \\ 0 & \text{if } N_{d_i, t_j} = 0 \end{cases}$$

where $w(i, j) \geq 0$, $tf$ and $idf$ are called the term frequency and inverse term frequency, respectively, $N_{d_i, t_j}$ denotes the number the term $t_j$ occurs in the document $d_i$ (term frequency factor), $N_{t_j}$ denotes the number of documents in collection $C$ in which $t_j$ occurs at least once (document frequency of the term $t_j$ ) and $|C|$ is the number of the documents in collection $C$. In Eq. (1), the first clause is applied for the words occurring in the document, whereas, the second clause is applied for words that do not appear (i.e., $N_{d_i, t_j} = 0$), for which $w(i, j) = 0$ is set. The frequency of the document is logarithmically scaled. The formula: $\log_2 \frac{|C|}{N_{t_j}} = \log C - \log N_{t_j}$ offers full weight to words occurring in a single document. A word occurring in entire documents would receive weight equal to zero. Using this weighting scheme, user-expected higher frequency keywords for generation of association rules were selected in this study.
**Step 2-Determination of predictive regions in texts:** Following the strategy adopted by [12], predictive regions in texts are identified in this step. Each region is embedded into $n_f$ dimensional space. We have used convolution neural network (CNN) for automatic feature extraction for a given class, say injury. The regions are identified as the most predictive one towards the output class after this training. In this case, all reports are input as the training set to model in the test mode and the predictive regions are recorded. For all incident outcomes, this process is carried out. These predictive regions are helpful in building more logical rules.

**Step 3-Development of association rules using ARM:** After determining predictive regions in texts, ARM is used for extracting interesting relationships hidden in accident data sets. The relationships can be represented in the form of sets of frequent items or association rules. The rules are generated on the basis of the frequency of item set that exists alone or in sequence with others in database [1]. Generally, $A \rightarrow B$ is a standard representation of an association rule, where $A$ is the antecedent and $B$ is the consequent, which conveys that $B$ will occur given that $A$ has already occurred for the same item in a database with a minimum threshold value. In the present study, Apriori algorithm is used for ARM technique on both types of data i.e., text and categorical. Although, the algorithm being same used in two cases, the basic concept of item sets in the algorithm is different. Words are considered as items in text-based ARM and classes in each of the categorical attributes are regarded as items in categorical-based ARM.

For example, if we consider ARM used on texts, then, given a set of keywords $A = \{w_1, w_2, ..., w_n\}$ and a collection of indexed documents $D = \{d_1, d_2, ..., d_m\}$, where each document $d_i$ is a set of keywords such that $d_i \subseteq A$. Let $W_j$ be a set of keywords. A document $d_i$ is said to contain $W_i$ if and only if $W_i \subseteq d_i$. An association rule is an implication of the form $W_i \rightarrow W_j$ where $W_i \subset A$, $W_j \subset A$, and $W_i \cap W_j = \emptyset$. There are two important basic measures for association rules i.e., Support ($S$) and Confidence ($C$). The rule $W_i \rightarrow W_j$ has support $S$ in the collection of documents $D$ if $S\%$ of documents in $D$ contain $W_i \cup W_j$. Then, Support is calculated by the following Eq. (2).

$$\text{Support}(W_i, W_j) = \frac{\text{Support count of } (W_i, W_j)}{\text{Total number of documents } D} \tag{2}$$

The rule $W_i \rightarrow W_j$ holds in the collection of documents $D$ with confidence $C$ if among those documents that contain $W_i$, $C\%$ of them contain $W_j$ also. We can compute confidence score using Eq. (3).

$$\text{Confidence}(W_i, W_j) = \frac{\text{Support}(W_i, W_j)}{\text{Support}(W_i)} \tag{3}$$

Basically, ARM is performed in the following two steps: (i) generate all the keyword combinations or sets, called as frequency sets, whose Support values are greater than the user specified minimum Support, and (ii) use the identified frequent keyword sets to generate the rules that satisfy a user specified minimum
confidence. The frequent keywords generation requires more effort and the rule
generation is straightforward. The \textit{Apriori} algorithm is used on the top 25% of
the keywords associated with higher \textit{TF-IDF} matrix but only the top 25% of
the keywords. The overall process of ARM on text data is depicted in Fig. 2. The
extracted association rules can be visualized in textual format or tables, or in
graphical format. In this study, it is designed to visualize the extracted associa-
tion rules in tabular format for a incident event leading to an incident outcomes
(i.e., injury, near miss, and property damage) in a particular department.

4 Results and Discussions

In this section, the rules obtained from both text and categorical data using
ARM are discussed below in brief.

4.1 Extraction of the Best Item-Set Rules from Textual Data

In this stage, ARM technique has been applied on accident text data i.e., incident
narratives. \textit{Apriori} algorithm has been deployed for this approach. A set of 23
rules has been extracted for all three incident outcome cases. Of which, nine rules
of injury are shown in Table 4. It can be observed that, in some cases, the rules
obtained from the text-based ARM form the common basic event as obtained
from fault tree analysis. For example, due to the occurrence of injury in the
department N2, some of the related bag of words are identified and consequently
rules using text-based ARM approach are generated like (left, operate, oven) \(\rightarrow\)
(STF). From this rule, it can be inferred that during the operation in oven in
the department N2, STF related causes resulting injuries are often happened.
In similar veins, from another rule (in Table 4), identified as (side, slip, toilet)
→ (STF), it can be interpreted that number of slipping incidents causing injury are observed more near a particular region like toilet area. It is noteworthy to mention that this important factor has been identified by text-based ARM whereas text mining analysis could hardly figure this out. Therefore, considering text-based ARM with text mining process can be an effective way of analysis to utilize the complete information within the body of unstructured text. In our study, text-based rules are obtained after meeting basic three criteria: (i) minimum threshold for Support and Confidence are set to be 0.002 and 0.01, respectively; (ii) the best rules are selected based on the highest Lift value; and (iii) the best rule are considered only to be a four-item or more than four-item set rules.

4.2 Extraction of the Best Item-Set Rules from Categorical Data

After extracting rules from text data, rules are extracted from categorical data. The combination/association of the factors in the form of rules is essential, which leads to accident scenarios. In order to find out the inherent rules, ARM has been used. Minimum value or threshold for Support \( (S) \), Confidence \( (C) \) is required to be set for extracting the rules. However, by changing the threshold, number of rules can be varied as per users’ requirement. Although there is no established condition mentioning the adoption of the thresholds, selection of those values depends on the data points used for the study, and usefulness of the strong rules. However, the minimum value for \( S \) and \( C \) each varies from 0.1% to 4%, and 1% to 34%, respectively [19, 41]. Another important parameter in ARM is lift \( (L) \) which is usually taken as greater than 1 or 2 [19]. In the present study, for injury, near miss, and property damage, the threshold values for \( S, C, \) and \( L \) are taken as 0.1%, 1.0%, and 1.0, respectively. The rules having \( S, C, \) and \( L \) values greater than threshold are selected finally, and others are discarded. Only the best rules (i.e., strong and interesting rules with higher Lift) for 3, 4, and 5-itemsets for injury, near miss, and property damage for each of the departments have been selected and shown in Tables 1, 2 and 3.

From Fig. 3(a), the most important three-item rule for injury in N1 department is found to be (incident type- IT1, observation type- OT3) → (incident category-injury) in N1 \( (S = 0.60\%, C = 9.50\%, L = 3.66) \) which describes that injury occurs mostly in N1 department due to behavioral related issue and unsafe acts by others. This finding supports the study of [24] in mining industry. Similarly, in case of four-item rule for the same case considered, the most important one having lift \( L = 12.80 \) is (working condition-W2 + machine condition-M3 + primary cause-R) → (incident category-injury) in N1 \( (S = 0.10\%, C = 33.30\%) \) which implies that rail is a reason for occurrence of injury in N1 department while the person is working alone and machine is running. Similarly, in case of five-item rule for the same case discussed, the most important rule having \( L = 1.51 \) is (SOP required-SPR3 + SOP available-SPAv1 + SOP compliance-SPC1 + SOP adequacy-SPA2) → (incident category-injury) in N1 department \( (S = 0.20\%, C = 3.90\%) \) which signifies that in N1 department injury has taken place for those jobs not following any SOP.
Fig. 3. Best association rules for (a) injury; (b) near-miss; (c) property damage in N1 department; and (d) attributes considered for generating three, four, and five item set rules in a department. Note: (IT = Incident types; OT = Observation types; PC = Primary causes; WC = Working conditions; MC = Machine conditions; SPR = SOP requirements; SPAv = SOP available; SOC = SOP compliance; SPA = SOP adequate).

From Fig. 3(b), the most important three-item set rule for near miss in N1 department is found to be (incident type- IT2 + observation type- OT3) → (incident category-near miss) \((S = 0.20\%, C = 11.10\%, L = 2.92)\) which narrates that near miss in N1 department occurs mostly due to process related issue as well as unsafe act by others. This result also supports the study of [41] on problems related to safety in a steel plant. Similarly, in case of four-item rule for the same case considered, the most important one having lift \(L = 15.76\) is (working condition-W2 + machine condition-M3 + primary cause-HM) → (incident category-near miss) in N1 \((S = 0.30\%, C = 60.00\%)\) which describes that hot metal is a reason for occurrence of near miss in N1 department while the person is working alone and machine is running. This rule supports the recent study of [38] in process industry. Similarly, among five-item set rules for the near miss, the best rule having lift \(L = 2.92\) is (SOP required-SPR3 + SOP available-SPAv1 + SOP compliance-SPC2 + SOP adequacy-SPA2) → (incident category-near miss) in N1 department \((S = 0.50\%, C = 11.10\%)\) which signifies that near miss in N1 department has happened for those jobs not following any SOP. In fact, this condition has been found to be prevailed in case of derailments after discussion with safety expert. Hence, SOP should be followed for every job in the industry.

Similarly, from Fig. 3(c), the most important three-item rule for property damage is observed as (incident type- IT2 + observation type- OT2) → (incident category-property damage) in N1 \((S = 0.20\%, C = 7.70\%, L = 4.52)\) which explains that property damage in N1 department occurs mostly due to process related issues and the presence of unsafe acts and unsafe conditions.
Table 1. Association rules for Injury cases in each department (8 departments, $8 \times 3 = 24$ best rules).

| Department | Association Rules (X $\rightarrow$ Y) | Support (S %) | Confidence (C %) | Lift (L) |
|------------|--------------------------------------|---------------|-----------------|----------|
| N1         | IT1 & OT3 $\rightarrow$ IN1          | 0.6           | 9.5             | 3.66     |
|            | M3 & W2 & R $\rightarrow$ IN1        | 0.1           | 33.3            | 12.8     |
|            | SPR3 & SPAv1 & SPC1 & SPA2 $\rightarrow$ IN1 | 0.2 | 3.9             | 1.51     |
| N2         | IT1 & OT3 $\rightarrow$ IN2          | 1             | 15.9            | 2.06     |
|            | M1 & W3 & STF $\rightarrow$ IN2      | 0.1           | 100             | 12.96    |
|            | SPR2 & SPAv2 & SPC3 & SPA3 $\rightarrow$ IN2 | 0.8 | 12.7            | 1.65     |
| N3         | IT1 & OT3 $\rightarrow$ IN3          | 0.8           | 12.7            | 5.51     |
|            | M3 & W1 & HP $\rightarrow$ IN3       | 0.2           | 40              | 17.36    |
|            | SPR3 & SPAv1 & SPC2 & SPA2 $\rightarrow$ IN3 | 0.2 | 4.4             | 1.93     |
| N4         | IT2 & OT1 $\rightarrow$ IN4          | 0.6           | 8.6             | 1.48     |
|            | M1 & W1 & SI $\rightarrow$ IN4       | 0.1           | 1               | 17.21    |
|            | SPR2 & SPAv2 & SPC3 & SPA3 $\rightarrow$ IN4 | 1.3 | 20.6            | 3.55     |
| N5         | IT2 & OT3 $\rightarrow$ IN5          | 0.1           | 5.6             | 2.41     |
|            | M1 & W2 & EMD $\rightarrow$ IN5      | 0.1           | 50              | 21.7     |
|            | SPR2 & SPAv2 & SPC3 & SPA3 $\rightarrow$ IN5 | 0.3 | 4.8             | 2.07     |
| N9         | IT1 & OT3 $\rightarrow$ IN9          | 0.4           | 6.3             | 1.71     |
|            | M1 & W2 & WAH $\rightarrow$ IN9      | 0.1           | 100             | 26.97    |
|            | SPR2 & SPAv2 & SPC3 & SPA3 $\rightarrow$ IN9 | 0.4 | 6.3             | 1.71     |
| N7         | IT1 & OT4 $\rightarrow$ IN7          | 3             | 11.7            | 1.77     |
|            | M1 & W3 & MA $\rightarrow$ IN7       | 0.1           | 100             | 15.12    |
|            | SPR2 & SPAv2 & SPC3 & SPA3 $\rightarrow$ IN7 | 1   | 15.9            | 2.4      |
| N8         | IT1 & OT3 $\rightarrow$ IN8          | 1.3           | 20.6            | 4.58     |
|            | M1 & W2 & OI $\rightarrow$ IN8       | 0.1           | 1               | 22.18    |
|            | SPR2 & SPAv2 & SPC3 & SPA3 $\rightarrow$ IN8 | 0.8 | 12.7            | 2.82     |

This also supports the study of [41] on safety associated incidents in the steel plant. Similarly, in case of four-item rule for the same case considered, the most important one having lift $L = 29.35$ is (working condition-W2 + machine condition-M1 + primary cause-PI) $\rightarrow$ (incident category-property damage) in N1 ($S = 0.10\%, C = 50.00\%)$ which reveals that process incident is also a reason for property damage in N1 department while person is working alone. The study on petroleum oil storage terminals by [36] also supports the above finding. Similarly, in case of five-item rule for the same case discussed, the most important rule having $L = 2.609$ is (SOP required-SPR3 + SOP available-SPAv1 + SOP compliance-SPC2 + SOP adequacy-SPA2) $\rightarrow$ (incident category-property damage) in N1 department ($S = 5.00\%, C = 4.40\%)$ which signifies that property damage in N1 department happened for those jobs not following any SOP. Figure 3(d) shows the overall rule generation structure from the data set. Three, four, and five item set rules have been generated for I, N, PD considering the attributes including IT & OT, PC, WC & MC, and SOPA, SOPC, SOPAv & SOPR, respectively (refer to Fig.3(d)).
Table 2. Association rules for near miss cases in each department (7 departments, \((7 \times 3 = 21\) best rules).

| Department | Association Rules \((X \rightarrow Y)\) | Support (S %) | Confidence (C %) | Lift (L) |
|------------|-------------------------------------|--------------|-----------------|---------|
| N1         | IT2 & OT3 \(\rightarrow\) NN1       | 0.2          | 11.1            | 2.92    |
|            | M3 & W2 & HM \(\rightarrow\) NN1    | 0.3          | 60              | 15.76   |
|            | SPR3 & SPAv1 & SPC2 & SPA2 \(\rightarrow\) NN1 | 0.5 | 11.1           | 2.92    |
| N2         | IT2 & OT1 \(\rightarrow\) NN2       | 1            | 14.3            | 1.62    |
|            | M2 & W3 & PI \(\rightarrow\) NN2    | 0.2          | 100             | 11.34   |
|            | SPR3 & SPAv1 & SPC3 & SPA3 \(\rightarrow\) NN2 | 0.1 | 100            | 11.34   |
| N4         | IT2 & OT3 \(\rightarrow\) NN4       | 0.3          | 16.7            | 7.23    |
|            | M1 & W2 & EI \(\rightarrow\) NN4    | 0.1          | 100             | 43.39   |
|            | SPR1 & SPAv2 & SPC3 & SPA3 \(\rightarrow\) NN4 | 0.7 | 15.2           | 6.6     |
| N5         | IT1 & OT1 \(\rightarrow\) NN5       | 2            | 6.1             | 2.02    |
|            | M2 & W1 & RO \(\rightarrow\) NN5    | 0.1          | 100             | 33.27   |
|            | SPR3 & SPAv1 & SPC2 & SPA1 \(\rightarrow\) NN5 | 2.3 | 10.6           | 3.53    |
| N9         | IT2 & OT2 \(\rightarrow\) NN9       | 0.3          | 11.5            | 2.13    |
|            | M2 & W3 & GL \(\rightarrow\) NN9    | 0.1          | 100             | 18.48   |
|            | SPR3 & SPAv1 & SPC1 & SPA2 \(\rightarrow\) NN9 | 0.7 | 13.7           | 2.54    |
| N7         | IT1 & OT2 \(\rightarrow\) NN7       | 0.9          | 10.3            | 2.07    |
|            | M2 & W1 & S \(\rightarrow\) NN7     | 0.1          | 100             | 20      |
|            | SPR3 & SPAv1 & SPC2 & SPA2 \(\rightarrow\) NN7 | 0.1 | 100            | 19.96   |
| N8         | IT2 & OT2 \(\rightarrow\) NN8       | 0.2          | 7.7             | 2.95    |
|            | M2 & W1 & EI \(\rightarrow\) NN8    | 0.1          | 100             | 38.39   |
|            | SPR3 & SPAv1 & SPC2 & SPA3 \(\rightarrow\) NN8 | 0.3 | 6.7            | 2.56    |

Table 3. Association rules for property damage cases in each department (8 departments, \((8 \times 3 = 24\) best rules).

| Department | Association Rules \((X \rightarrow Y)\) | Support (S %) | Confidence (C %) | Lift (L) |
|------------|-------------------------------------|--------------|-----------------|---------|
| N1         | IT2 & OT1 \(\rightarrow\) PN1       | 0.2          | 7.7             | 4.52    |
|            | M1 & W2 & PI \(\rightarrow\) PN1    | 0.1          | 50              | 29.35   |
|            | SPR3 & SPAv1 & SPC2 & SPA2 \(\rightarrow\) PN1 | 0.2 | 4.4            | 2.61    |
| N2         | IT2 & OT2 \(\rightarrow\) PN2       | 0.4          | 15.4            | 2.9     |
|            | M1 & W3 & EMD \(\rightarrow\) PN2   | 0.1          | 100             | 18.83   |
|            | SPR3 & SPAv1 & SPC1 & SPA2 \(\rightarrow\) PN2 | 1.6 | 12.6           | 2.37    |
| N4         | IT2 & OT4 \(\rightarrow\) PN4       | 0.8          | 5.4             | 3.17    |
|            | M1 & W3 & SI \(\rightarrow\) PN4    | 0.1          | 100             | 58.71   |
|            | SPR1 & SPAv2 & SPC3 & SPA3 \(\rightarrow\) PN4 | 0.4 | 8.7            | 5.11    |
| N5         | IT1 & OT1 \(\rightarrow\) PN5       | 2.2          | 6.7             | 1.75    |
|            | M2 & W3 & EMD \(\rightarrow\) PN5   | 0.1          | 100             | 26.26   |
|            | SPR3 & SPAv1 & SPC2 & SPA1 \(\rightarrow\) PN5 | 2.7 | 12.4           | 3.27    |
| N9         | IT2 & OT2 \(\rightarrow\) PN9       | 0.5          | 19.2            | 5.48    |
|            | M2 & W1 & EMD \(\rightarrow\) PN9   | 0.2          | 50              | 14.26   |
|            | SPR1 & SPAv2 & SPC3 & SPA3 \(\rightarrow\) PN9 | 0.4 | 8.7            | 2.48    |
| N6         | IT2 & OT1 \(\rightarrow\) PN6       | 0.1          | 1.4             | 1.43    |
|            | M1 & W1 & PI \(\rightarrow\) PN6    | 0.1          | 33.3            | 33.27   |
|            | SPR3 & SPAv1 & SPC1 & SPA2 \(\rightarrow\) PN6 | 0.1 | 2             | 1.96    |
| N7         | IT1 & OT1 \(\rightarrow\) PN7       | 8.8          | 26.7            | 1.99    |
|            | M2 & W2 & R \(\rightarrow\) PN7     | 0.1          | 100             | 7.45    |
|            | SPR3 & SPAv1 & SPC2 & SPA1 \(\rightarrow\) PN7 | 4.3 | 19.8           | 1.48    |
| N8         | IT2 & OT3 \(\rightarrow\) PN8       | 0.2          | 11.1            | 3.7     |
|            | M2 & W2 & S \(\rightarrow\) PN8     | 0.1          | 100             | 33.27   |
|            | SPR1 & SPAv2 & SPC3 & SPA3 \(\rightarrow\) PN8 | 0.5 | 10.9           | 3.62    |
Table 4. Key factors with corresponding rules from the proposed model for injury across all the departments.

| Incident Category | Frequently occurring primary cause | Relevant rule from Text based ARM (Considering only the more than 4-item set rules for meaningful interpretation) | Best Rule from association rule mining (3-item rules) | Best Rule from association rule mining (4-item rules) | Best rule from association rule mining (5-item rules) |
|------------------|-----------------------------------|-----------------------------------------------------------------------------------------------------------------|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|
| IN1               | (bike, come, duty) ⇒ [IZ11 RI]   | Due to unsafe act by others and behavioral related problem; Rail is a reason when worker is working alone          | Inadequate SOP                                    |                                                                                       | Available, inadequate SOP                           |
|                  | (come, duty, injury) ⇒ [IZ2 RI]  |                                                                                                                 |                                                                                       |                                                                                       |                                                                                       |
| IN2               | (side, slip, tasks) ⇒ [IZ2 STF]  | Due to unsafe act by others and behavioral related problem; STF is a reason in non-working condition               | Inadequate SOP                                    |                                                                                       |                                                                                       |
|                  | (left, operate, oven) ⇒ [IZ2 STF]|                                                                                                                 |                                                                                       |                                                                                       |                                                                                       |
| IN3               | (cycle, mat, ride) ⇒ [IZ1 RI]    | Due to unsafe act by others and behavioral related problem; Hydraulic/Pneumatic is a reason when persons are working in group | Available, inadequate SOP                         |                                                                                       |                                                                                       |
|                  | (after, complete, duty, home) ⇒ [IZ1 RI] |                                                                                                               |                                                                                       |                                                                                       |                                                                                       |
| IN4               | (employee, slip, space) ⇒ [IZ4 STF] | Due to unsafe act by person and process behavioral related problem; Structural integrity is a reason when people are working in group | Unavailable SOP                                  |                                                                                       |                                                                                       |
|                  | (helper, proper, work) ⇒ [IZ4 STF]|                                                                                                                 |                                                                                       |                                                                                       |                                                                                       |
| IN5               | [injury] ⇒ [IZ2 STF]              | Due to unsafe act by others and process equipment machinery damage; person is working alone                        | Unavailable SOP                                  |                                                                                       |                                                                                       |
|                  |                                                                                       | Working at height is a reason when person is working alone                                                      |                                                                                       |                                                                                       |                                                                                       |
| IN7               | (empty, got, injury, loss) ⇒ [IZ7 STF]| Due to unsafe act by others and behavioral related problem; Equipment machinery damage is a reason when person is working alone | Unavailable SOP                                  |                                                                                       |                                                                                       |
|                  | (couple, got, wagon) ⇒ [IZ7 STF]|                                                                                                                 |                                                                                       |                                                                                       |                                                                                       |
| IN8               | [cover, ground, release] ⇒ [IZ2 MH] | Due to unsafe condition and behavioral related problem; Medical alignment is a reason                             | Unavailable SOP                                  |                                                                                       |                                                                                       |
|                  | (aid, clean, leg) ⇒ [IZ2 MH]|                                                                                                                 |                                                                                       |                                                                                       |                                                                                       |
| IN9               | (grain, inba, inspect, tank, while) ⇒ [IZ9 STF]| Due to unsafe act by others and behavioral related problem; Occupational illness is a reason when person is working alone | Unavailable SOP                                  |                                                                                       |                                                                                       |
|                  | (inspect, tank, while) ⇒ [IZ9 STF]|                                                                                                                 |                                                                                       |                                                                                       |                                                                                       |

5 Conclusions

In the present study, a hybrid model encompassing text mining-based ARM approach has been proposed to predict the incident outcomes i.e., injury, near miss, and property damage and to investigate the accidents through rule-based analysis in various departments in the steel plant. The main aim of this study is to use categorical and textual data together in building a prediction model which could predict the incident outcomes in different departments of the plant and finding association rules among various factors leading to incidents. In order to extract information from the unstructured text field, text-based ARM approach has been applied to uncover the hidden patterns in terms of rules or associations among the factors which might not be captured by human judgment. In addition, ARM on categorical data explores some interesting and useful findings. For examples, in N1 department, injury mostly occurs due to behavioral related issue and unsafe act. Rail is also found to be a reason for injury in the same department while person is working alone and machine is in running condition. Analyses also indicate that ignoring or not following SOP also leads to injury in the N1 department.
Like other research, the study has also some limitations. First, data pre-processing including data cleaning, standardization, consistency check could be taken care of efficiently by well-established algorithms to increase the accuracy of the model. Second, higher the data size, higher will be the generalization ability of the model. But, in this study, very limited data has been used to build the model. Another limitation of our study is that while generating association rules from text, sometimes very limited number of rules (sometimes even no rules) are generated with minimum support and confidence value which is practically very difficult for interpretation. As future aspects of this study, survey can be done for the further verification as well as validation of our proposed model by expert opinions from particular domain. Some unsupervised learning algorithms like clustering could be implemented. Moreover, other algorithms of association rule mining approaches such as predictive Apriori, Equivalence Class Transformation (ECLAT) algorithm could be implemented for better investigation of the problem. Decision tree modeling could also be another kind of effective approach for further investigation of occupational accident analysis in industrial domain. In addition, this paper only deals with the best rules derived from ARM, avoiding taking other rules. Future study could further investigate some high-quality item set rules finding from the data set. New algorithm for text-based ARM could be developed in order to find out more hidden interactions of factors with the reduction of human effort. Furthermore, ontology-based text mining could be incorporated at the earlier stage of the study. On top of that, though the scope of the present study is limited to steel industry, it can be implemented in other application domains also such as mining, manufacturing, construction, aviation etc.

Acknowledgment. The work is funded by UAY project, GOI (Project Code: IITKGP_022). We acknowledge the Centre of Excellence in Safety Engineering and Analytics (CoE-SEA) (www.iitkgp.ac.in/department/SE) and Safety Analytics & Virtual Reality (SAVR) Laboratory (www.savr.iitkgp.ac.in) of Department of Industrial & Systems Engineering, IIT Kharagpur for experimental/computational and research facilities for this work. We would like to thank the management of the plant for providing relevant data and their support and cooperation during the study.

References

1. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), vol. 1215, pp. 487–499 (1994)
2. Amiri, M., Ardeshir, A., Fazel Zarandi, M.H., Soltanaghaei, E.: Pattern extraction for high-risk accidents in the construction industry: a data-mining approach. Int. J. Inj. Control Saf. Promot. 23(3), 264–276 (2016)
3. Basso, B., Carpegna, C., Dibitonto, C., Gaido, G., Robotto, A., Zonato, C.: Reviewing the safety management system by incident investigation and performance indicators. J. Loss Prev. Process Ind. 17(3), 225–231 (2004)
4. Bertke, S., Meyers, A., Wurzelbacher, S., Bell, J., Lampl, M., Robins, D.: Development and evaluation of a Na"ive Bayesian model for coding causation of workers’ compensation claims. J. Saf. Res. 43(5), 327–332 (2012)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
6. Brown, D.E.: Text mining the contributors to rail accidents. IEEE Trans. Intel. Transp. Syst. 17(2), 346–355 (2016)
7. Bunn, T.L., Slavova, S., Hall, L.: Narrative text analysis of Kentucky tractor fatality reports. Accid. Anal. Prev. 40(2), 419–425 (2008)
8. Cheng, C.W., Lin, C.C., Leu, S.S.: Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction industry. Saf. Sci. 48(4), 436–444 (2010)
9. Chi, C.F., Lin, S.Z., Dewi, R.S.: Graphical fault tree analysis for fatal falls in the construction industry. Accid. Anal. Prev. 72, 359–369 (2014)
10. Fu, Z., Wu, X., Wang, Q., Ren, K.: Enabling central keyword-based semantic extension search over encrypted outsourced data. IEEE Trans. Inf. Forensics Secur. 12(12), 2986–2997 (2017)
11. ILO: Protecting Workplace Safety and Health in Difficult Economic Times. The Effect of the Financial Crisis and Economic Recession on Occupational Safety and Health. Programme on Safety and Health at Work and the Environment (Safe-Work). Seiji Machida Director. Technical Report (2013)
12. Johnson, R., Zhang, T.: Effective use of word order for text categorization with convolutional neural networks. arXiv preprint arXiv:1412.1058 (2014)
13. Kumar, S., Toshniwal, D.: A data mining framework to analyze road accident data. J. Big Data 2(1), 26 (2015)
14. Lehto, M., Marucci-Wellman, H., Corns, H.: Bayesian methods: a useful tool for classifying injury narratives into cause groups. Inj. Prev. 15(4), 259–265 (2009)
15. Li, C., Qin, J., Li, J., Hou, Q.: The accident early warning system for iron and steel enterprises based on combination weighting and grey prediction model GM (1, 1). Saf. Sci. 89, 19–27 (2016)
16. Liao, C.W., Perng, Y.H.: Data mining for occupational injuries in the Taiwan construction industry. Saf. Sci. 46(7), 1091–1102 (2008)
17. Liao, C.W., Perng, Y.H., Chiang, T.L.: Discovery of unapparent association rules based on extracted probability. Decis. Support Syst. 47(4), 354–363 (2009)
18. Mirabadi, A., Sharifian, S.: Application of association rules in Iranian railways (rai) accident data analysis. Saf. Sci. 48(10), 1427–1435 (2010)
19. Montella, A., Aria, M., D’Ambrosio, A., Mauriello, F.: Data-mining techniques for exploratory analysis of pedestrian crashes. Transp. Res. Rec. J. Transp. Res. Board 2237, 107–116 (2011)
20. Nenonen, N.: Analysing factors related to slipping, stumbling, and falling accidents at work: application of data mining methods to finnish occupational accidents and diseases statistics database. Appl. Ergon. 44(2), 215–224 (2013)
21. Nordlöf, H., Wiittavaara, B., Winblad, U., Wijk, K., Westerling, R.: Safety culture and reasons for risk-taking at a large steel-manufacturing company: investigating the worker perspective. Saf. Sci. 73, 126–135 (2015)
22. Pande, A., Abdel-Aty, M.: A computing approach using probabilistic neural networks for instantaneous appraisal of rear-end crash risk. Comput. Aided Civ. Infrastructure. Eng. 23(7), 549–559 (2008)
23. Patel, M.D., Rose, K.M., Owens, C.R., Bang, H., Kaufman, J.S.: Performance of automated and manual coding systems for occupational data: a case study of historical records. Am. J. Ind. Med. 55(3), 228–231 (2012)
24. Paul, P., Maiti, J.: The role of behavioral factors on safety management in underground mines. Saf. Sci. 45(4), 449–471 (2007)
25. Pramanik, A., Sarkar, S., Maiti, J.: Oil spill detection using image processing technique: an occupational safety perspective of a steel plant. In: Emerging Technologies in Data Mining and Information Security, vol. 814, pp. 247–257. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1501-5_21

26. Sanchez-Pi, N., Martí, L., Garcia, A.C.B.: Text classification techniques in oil industry applications. In: International Joint Conference SOCO’13-CISIS’13-ICEUTE’13, pp. 211–220. Springer (2014). https://doi.org/10.1007/978-3-319-01854-6_22

27. Sanchez-Pi, N., Martí, L., Garcia, A.C.B.: Improving ontology-based text classification: an occupational health and security application. J. Appl. Log. 17, 48–58 (2016)

28. Sarkar, S., Verma, A., Maiti, J.: Prediction of occupational incidents using proactive and reactive data: a data mining approach. In: Maiti, J., Ray, P.K. (eds.) Industrial Safety Management. MAC, pp. 65–79. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-6328-2_6

29. Sarkar, S., Baidya, S., Maiti, J.: Application of rough set theory in accident analysis at work: a case study. ICRCICN 2017, 245–250 (2017)

30. Sarkar, S., Chain, M., Nayak, S., Maiti, J.: Decision support system for prediction of occupational accident: a case study from a steel plant. In: Emerging Technologies in Data Mining and Information Security, vol. 813, pp. 787–796. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1498-8_69

31. Sarkar, S., Ejaz, N., Maiti, J.: Application of hybrid clustering technique for pattern extraction of accident at work: a case study of a steel industry. In: 2018 4th International Conference on Recent Advances in Information Technology (RAIT), pp. 1–6. IEEE, IIT Dhanbad (2018)

32. Sarkar, S., Kumar, A., Mohanpuria, S.K., Maiti, J.: Application of Bayesian network model in explaining occupational accidents in a steel industry. In: ICRCICN 2017, pp. 337–342. IEEE (2017)

33. Sarkar, S., Lakha, V., Ansari, I., Maiti, J.: Supplier selection in uncertain environment: a fuzzy MCDM approach. In: ICIC2 - 2016, pp. 257–266. Springer (2017)

34. Sarkar, S., Pateshwari, V., Maiti, J.: Predictive model for incident occurrences in steel plant in India. ICCCNT 2017, 1–5 (2017)

35. Sarkar, S., Raj, R., Sammangi, V., Maiti, J., Pratihar, D.: An optimization-based decision tree approach for predicting slip-trip-fall accidents at work. Saf. Sci. 118, 57–69 (2019)

36. Sharma, R.K., Gurjar, B.R., Singhal, A.V., Wate, S.R., Ghuge, S.P., Agrawal, R.: Automation of emergency response for petroleum oil storage terminals. Saf. Sci. 72, 262–273 (2015)

37. Singh, K., Raj, N., Sahu, S., Behera, R., Sarkar, S., Maiti, J.: Modelling safety of gantry crane operations using petri nets. Int. J. Inj. Control Saf. Promot. 24(1), 32-43 (2015)

38. Stefana, E., Marciano, F., Alberti, M.: Qualitative risk assessment of a dual fuel (LNG-diesel) system for heavy-duty trucks. J. Loss Prev. Process Ind. 39, 39–58 (2016)

39. Taylor, J.A., Lacovara, A.V., Smith, G.S., Pandian, R., Lehto, M.: Near-miss narratives from the fire service: a Bayesian analysis. Accid. Anal. Prev. 62, 119–129 (2014)

40. Verma, A., Chatterjee, S., Sarkar, S., Maiti, J.: Data-driven mapping between proactive and reactive measures of occupational safety performance. In: Maiti, J., Ray, P.K. (eds.) Industrial Safety Management. MAC, pp. 53–63. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-6328-2_5
41. Verma, A., Khan, S.D., Maiti, J., Krishna, O.: Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. Saf. Sci. 70, 89–98 (2014)
42. Zheng, W., Shuai, J., Shan, K.: The energy source based job safety analysis and application in the project. Saf. Sci. 93, 9–15 (2017)