The XXXVI[th] Annual International Occupational Ergonomics and Safety Conference
Denver, Colorado
August 5-6, 2024

Development of Accident Path/chain using Semi-supervised Topic Modelling: A Novel Methodology with Application in Petrochemical Industry

Jhareswar Maiti[1], and Satyajeet Sahoo[2]

[1]Department of Industrial Engineering, Indian Institute of Technology Kharagpur, Kharagpur-721302, West Bengal, India
[1]CoE in Safety Engineering and Analytics, Indian Institute of Technology Kharagpur, Kharagpur-721302, West Bengal, India
[2]Department of Industrial Engineering, Indian Institute of Technology Kharagpur, Kharagpur-721302, West Bengal, India

Corresponding author's Email: jmaiti@iem.iitkgp.ac.in

Author Note: Prof. Jhareswar Maiti is the Head of the Department of Industrial and Systems Engineering (ISE) and Chairman of Centre of Excellence in Safety Engineering and Analytics (CoE-SEA), IIT Kharagpur. His research interests include safety, reliability and risk modelling using industry 4.0 techniques like analytics, VR and machine learning. Mr. Satyajeet Sahoo is currently pursuing PhD under his supervision in the domain of safety engineering and analytics.

Abstract: Traditionally, process industries have relied on engineering first-principles based accident models to perform Fault Detection and Diagnosis (FDD) and predict accidents. However, with the advent of Industry 4.0, there is increasing adoption of data-mining models to identify patterns from data and use the patterns to perform fault diagnosis and accident modelling. Safety documents, in the form of visit/observation reports, audits and incident investigation reports contain considerable information on faults and possible failure events in process systems and use of text mining models can reveal insights on various latent faults/root causes and how their interactions result in propagation of hazards into accidents. Here a novel semi-supervised text mining methodology is presented that enables the user to automatically identify various Risk Control System (RCS) failures in process plants. Using user-defined RCS-relevant anchor words as inputs, the model performs topic modelling on incident descriptions to identify the root causes of accidents in form of latent topics representing RCS failures. Then association rule mining is applied on the RCS failure item-sets to develop major accident chains. This methodology can not only predict subsequent RCS failure events and possible accidents, but can also provide insights on failure proneness of RCS across organizational units. Based on the failure analysis, this methodology can recommend design and deployment of safety barriers in an automated manner. The methodology was applied on incident investigation reports at a petrochemicals plant and 15 accident chains representing sequence of failure events causing propagation of hazards into accidents were identified. By developing appropriate industry-specific anchor words, the methodology can be applied in other industries and important root causes and sequence of failure events resulting in accidents can be identified.

Keywords: Accident Path, Chain of Events, Risk Control Systems, Correlation Explanation, Anchor Words

1. Introduction

Risk management of process systems and operations consists of two aspects: a) Detection and diagnosis of the possible faults inherent in a system, and b) Modelling the behavior of the faults using failure and accident models to understand and predict how hazard propagates due to interplay of various faults and results in incidents. One of the prominent accident models is the accident path, which describes fault propagation by occurrence of a rocesss to use domain experts to identify the various accident paths in process systems. However, with increase in safety awareness, many safety programs and initiatives are being undertaken in process plants which generate a lot of text data in form of reports on which text mining tools can be applied to automate fault detection and accident modelling of process systems, develop accident paths, and predict future failure events and their impact. Here we propose a novel semi-supervised framework that produces accident paths by performing text mining of incident descriptions in safety reports using Correlation Explanation (CorEx) topic model. By developing a dictionary of prominent anchor words for various Risk Control Systems (RCS) applicable in various process systems and subsystems, and feeding the anchor words to CorEx model as guiding words, the various faults in form of topics corresponding to RCS failures that happened in the system/subsystem under consideration are identified. Then applying Apriori algorithm, various rules of association are mined between the faults and thus the chain of events is obtained. The framework is applied on corpus of incident reports at a prominent petrochemical plant in India and 13 major chains of events/accident paths are obtained.

ISBN: 9781938496622
DOI: https://doi.org/10.47461/ISOES.2024.Maiti

2. Literature Review

Risk assessment of process systems requires understanding the various types of hazards/faults inherent in the system and its constituent sub-systems and components using Fault Detection and Diagnosis (FDD), and identifying the causal relationships between the faults using accident models to understand their behavior and interactions and thus simulate scenarios of possible failure propagation paths resulting in accident scenarios (Khan and Abbasi, 1998). The prevailing practice in process industry for accident modelling is to use various qualitative and quantitative accident models developed by researchers and domain experts (Khan et al, 2015). Accident models like Chain of Events (Triplett et al, 2004), Domino theory (Heinrich, 1941), and Swiss Cheese model (Reason,1990) are extensively used in process safety to understand how cause-effect interactions and alignments between multiple faults influences characteristics of hazard propagation. Chain of events (also called accident path) model propagation of hazards from hazardous element to final incident via sequential occurrence of multiple faults/failure events, forming an event chain. This sequence of failure events is called accident-causing mechanism (ACM), where the initial fault that triggers the chain reaction is called the initiating event and subsequent events are called pivotal events (Singh and Maiti, 2020). The hazard propagation can be stopped by breaking the chain reaction by putting in place preventive and mitigating risk control systems (RCS). Nine RCS have been proposed in the accident path model-maintenance and inspection, procedures, competency of plant staff, instrument and alarms, permit to work, communication, plant design, safety barriers and energy isolation, and emergency preparedness.

Increasing awareness of hazards and the need for proactive safety management is resulting in real-time monitoring of processes, equipment and operations using sensors and cameras. Moreover, many safety programs and audits are increasingly being implemented in process plants. This is resulting in generation of huge volume of process parameters' data, images, videos, and safety reports consisting of description of incidents, near-misses, and observations during inspections. These reports are an important source of information to perform accident modelling and risk assessment using text mining models. A review of literature shows extensive use of supervised text mining models for accident modelling and risk analysis (Ekramipooya et al, 2023), fault identification in coal mines (Qiu et al., 2021), or identification of hazardous elements and accident-causing mechanisms at steel plants (Sahoo and Maiti, 2022). Supervised models require training data, and a process organization needs to invest significant time and resources in collecting, cleaning, pre-processing, and annotating such data. Hence there is great scope of developing unsupervised accident modelling frameworks. Application of unsupervised models for risk assessment and in literature reveals identification of faults and associated accident paths and RCS have been using K-mode clustering (Singh et al. 2019) or using Directed Acyclic Network (DAN) to identify probable chain of events and model propagation of failures and associated risks (Meng et al.). One important category of unsupervised models includes topic models, which are used to identify latent themes in text. Faults and failure events can be represented by topics in incident descriptions and can be identified and extracted using topic models. One of the most popular topic models is Latent Dirichlet Allocation (LDA) (Blei, 2012) which is seeing increasing adoption in text mining of reports for incident classification (Ertek and Kailas, 2021), identification of co-occurrence patterns of attributes (Roque et al., 2019), generation of temporal trends in themes (Robinson, 2019) etc. However, as topic models are unsupervised, it becomes difficult to identify and annotate topics from the keywords generated by LDA. Hence fault detection using topic models becomes difficult. Secondly, while topics can be identified, the causal relationships between topics, which is necessary to model the sequential occurrence of faults in chain of events, is difficult to be obtained without expert input. While topic modelling combined with association rule mining have been applied to identify chain of events/accident paths and fault trees to assess contractor risk at steel plants (Sahoo et al., 2024), a relative lack of literature is observed regarding application of topic models and other unsupervised techniques for accident modelling of process systems.

Here a novel ensembled framework is proposed that combines Correlation Explanation (CorEx), a semi-supervised topic model that obtains topics using a dictionary of anchor words (Gallagher et al., 2017), with association rule mining, to obtain chain of events/accident path. The CorEx model is an information-theoretic framework used to obtain maximally informative topics from the text data. A case study is presented here where CorEx model is applied on incident investigation reports of a petrochemical plant in India to identify critical faults as RCS failures and generate major accident paths prevalent in the plant. First, using inputs from domain experts, a dictionary of keywords is developed for each RCS failure. These keywords become the anchor words which are then fed to CorEx model to identify the topics that have maximal information w.r.t. the RCS keywords and topics corresponding to RCS failures are identified from each incident report. By obtaining the set of all RCS failures for each incident and applying Apriori algorithm, the strong rules of association are established between the faults and chain of events is obtained.

3. Preliminaries

3.1 CorEx

Correlation Explanation (CorEx), introduced by Gallagher et al. (2017), is a semi-supervised topic model that identifies relevant topics in text based on various topic "anchor words" developed using experts' feedback. Let a discrete random variable be represented by X. Let the observations of X be represented as x_i. The total correlation of a subset of X, represented by X_G is given by: $TC(X_G) = \sum_{i \in G} \quad E(X_i) - E(X_G)$, where E(X) and $E(X_G)$ represent the entropy of X and X_G, respectively.

In context topic models, X_G represents groups of related words and let Y represents the latent theme (topic) that needs to be extracted from a document. Considering topics $Y_1, \ldots Y_M$, CorEx tries to maximize the reduction of total conditional correlation, TC (X_G; $Y_1, \ldots, Y_M$) by maximizing the lower bound of $\sum_{j=1}^{m} \quad TC\left(X_{G_J}; Y_j\right)$, so that words dependencies in documents can be maximally explained. These words then represent the keywords of topics. For further details, readers may refer to Gallagher et al. (2017).

3.1 Accident Path

An accident path is a set of 5 tuples consisting of sources of hazards (Hazardous Element (HE)) and sequential chain consisting of the trigger event (Initiating Event (IE)) initiating sequential occurrences of subsequent events (Pivotal Events (PV)) resulting in hazard and failure propagation culminating in final incident causing harm (Threat) to human, property, or environment (Target). This hazard propagation can be stopped by placing preventive risk control systems to break the chain and mitigating risk control systems to mitigating impact of final incident. A schematic diagram shown in Fig. 1:

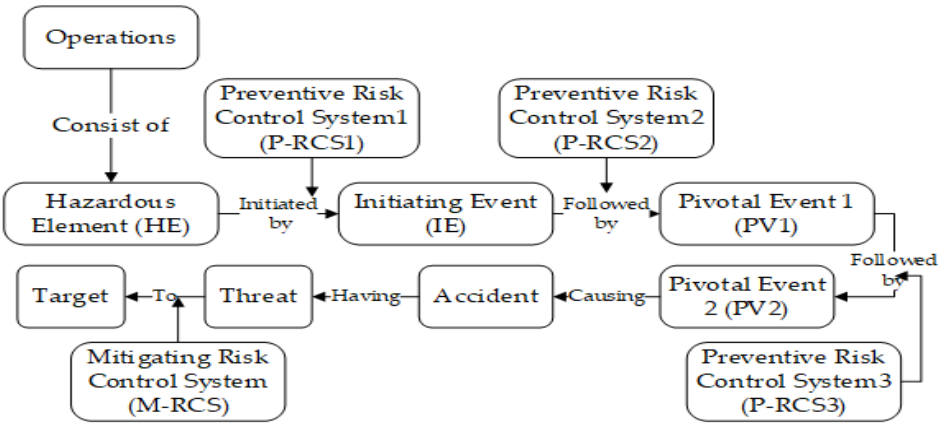

Fig. 1: Schematic diagram of Accident Path

4. Methodology

The framework consists of the following steps:

4.1 Text Pre-processing

Incident descriptions in the safety reports consists of unstructured text written in various styles and formats. Hence text pre-processing is carried out where punctuation marks are removed, characters converted to lowercase, stop words are removed, unique words (tokens) are extracted, and the tokens are converted to base lemma form using lemmatization. After standardization of tokens, they are converted into quantitative values using vector embeddings. Here in this study, we have used TF-IDF embeddings, which capture the relevancy of each token in the text. Thus, for each document in the dataset, a global vector of all unique tokens in the corpus is created, with the contribution of each token of the corpus to the document given by respective TF-IDF values. On this vector we then apply CorEx model.

4.2 Application of CorEx and obtaining item-sets of RCS failures

The next step involves application of Correlation Explanation (CorEx) to identify the various faults/failure events in an incident description. As per the accident path theory, a chain of events results in propagation of failure into an accident by sequential occurrence of events one after another in a domino fashion, and hazard propagation can be stopped by applying one of the risk control systems (RCS). Since there is an RCS for every failure event of a chain, conversely, we can say that each failure event can be described as failure of an RCS. Hence this model aims at identifying the faults/failure events in the incident descriptions as an RCS failure by first developing a dictionary of keywords corresponding to each RCS failure and then feeding this dictionary as anchor words to the CorEx topic model. On applying this semi-supervised technique on the incident reports, various topics having maximum correlation to the various RCS violation keywords are obtained. These topics then represent the faults/failure events in form of RCS failures. Hence for each incident report, we obtain sets of various possible RCS failure faults that occurred in the incident description.

4.3 Obtaining rules of association between the RCS failures and developing the chain of events

The final step in obtaining the chain of events involves establishing the causal relationships between the various RCS failures by establishing the rules of association between them. For this, the sets of RCS failures for the incident reports in the corpus that were obtained by applying anchor words based CorEx as considered as item-sets on which association rule mining is applied. By applying suitable levels of support and confidence, frequent item-sets are obtained and Apriori algorithm is applied to obtain the rules of association between the faults. These rules obtained are considered as cut-sets which are further pruned using principles of minimal cut-sets to obtain the rules that consist of all those faults that form minimal cut-sets for all the incidents described in the incident reports in the corpus. Thus, the final list of critical chain of events for the process system under consideration are obtained.

5. Results

A dictionary of anchor words was developed using expert inputs for various RCS, a sample shown in Table 1:

Table 1. Sample anchor-words for various RCS failures

Inspection & maintenance	Staff Competency	Procedures	Instruments and alarm	Communication	Permit to Work	Safety Barrier and Isolation	Emergency Arrangements
Apparatus Equipment Tool Breakdown Repair Defect Malfunction	Carelessness Negligence Absent-minded Employee Mistake Responsible	Rule Approach Manual Procedure Instructions Specifications	Detector Meter Alarm Gauge Instrument Sensor	Report Circular Permission Misunderstanding Miscommunication	Permit Authorization License Approval Grant	Barricade Handrail Scaffold Danger Earthing Activation	Ambulance Dispensary First aid Lost time Clinic Injury Emergency

Applying the anchor words infused CorEx to the incident descriptions reveals the various RCS failures present in the incident description. Table 2 gives the RCS failures obtained for sample of 5 incident reports from the petrochemical plant:

Table 2: RCS failures identified in incident reports (1= failure happened, 0= failure did not happen)

Incident Description	**ITPM**	**SC**	**SOP**	**PTW**	**PEI**	**EA**
injure person assign load truck pp warehouse loading leave leg get twist presence crack truck bed disguise tarpaulin cover cause injury	0	1	0	0	0	1
v bottom conical hydrocarbon leak arrest job carry workman scaffold platform height sufficient access clearance find job execution due exist cable tray pipeline support structure job location however bolt tighten job use spanner hammer carry scaffold platform hammering apply hammer spanner tighten bolt hammer get misbalance suddenly hand due hammer hit cable tray hammer fall leave index finger hold spanner concerned bolt get injure leave index finger	1	1	0	1	1	0
driver truck injuired fixing last batta right hand trap batta truck body	0	1	0	0	0	0
date time worker return work lunch middle road fall due false step get injury leave leg knee light swelling observe immediately apply pain reliever spray shift medical center	0	1	0	0	1	1

injury cause bee_sting bite electrical_connection street_light front_auxulary fire_station	0	0	0	0	0	1

Where ITPM is Inspection Testing and Maintenance, SC is Staff Competency, SOP is Standard Operating Procedure, PTW is Permit to Work, PEI is Barriers and Positive Energy Isolation and EA is Emergency Arrangements. Taking all available RCS failures in an incident description as an item-set, and applying association rule mining and Apriori algorithm on all such item-sets obtained, we obtain frequent item-sets and rules of association between the faults as given in Table 3:

Table 3: Rules of association between the RCS violations

Sl No	Antecedents	Consequents	Support	Confidence	Lift	Leverage	Convictio n	Zhang's metric
1	'SC'	' EA'	0.155	0.364	1.466	0.049	1.182	0.554
3	'ITPM'	' PEI'	0.147	0.543	2.334	0.084	1.679	0.784
4	' SC'	' PEI'	0.116	0.375	1.613	0.044	1.228	0.551
5	ITPM', ' SOP'	' SC'	0.101	0.813	2.620	0.062	3.680	0.706
6	ITPM', ' SOP'	' PEI'	0.093	0.750	3.225	0.064	3.070	0.788
7	'ITPM'	' PEI', ' SC'	0.093	0.343	2.949	0.061	1.345	0.907
8	'ITPM', ' PTW', ' SC'	' PEI'	0.070	0.692	2.977	0.046	2.494	0.739
9	' SOP'	'ITPM'	0.070	0.600	2.211	0.038	1.822	0.620
10	' SOP'	' SC'	0.070	0.600	1.935	0.034	1.725	0.547
11	'ITPM', ' SC'	' EA'	0.070	0.346	1.395	0.020	1.150	0.355
12	' SOP'	' PEI'	0.054	0.467	2.007	0.027	1.439	0.568
13	SOP', ' SC'	'ITPM'	0.039	1.000	3.686	0.028	inf	0.758
14	' SOP'	' PEI', ' SC'	0.039	0.333	2.867	0.025	1.326	0.737
15	' SC'	EA', ' SOP'	0.031	0.100	3.225	0.021	1.077	1.000
16	'SC'	' EA', ' PEI'	0.031	0.073	1.340	0.008	1.020	0.443

Finally, list of chain of events after identifying the minimum cut-sets and pruning redundant rules is given in Table 4:

Table 4: Final accident paths obtained for various incidents of the petrochemical plant

Sl No	**Initiating Event**	**Pivotal Event-1**	**Pivotal Event-2**	**Pivotal Event-3**
1	Competency Issues	Emergency Arrangements		
2	Competency Issues	Equipment Issues	Emergency Arrangements	
3	Competency Issues	Equipment Issues	Safety Barrier/isolation issues	
4	Competency Issues	Safety Barrier/isolation issues	Emergency Arrangements	
5	Equipment Issues	Emergency Arrangements		
6	Equipment Issues	Safety Barrier/isolation issues		
7	SOP & PTW	Competency Issues	Emergency Arrangements	
8	SOP & PTW	Competency Issues	Equipment Issues	Safety Barrier/isolation issues
9	SOP & PTW	Competency Issues	Equipment Issues	
10	SOP & PTW	Competency Issues	Safety Barrier/isolation issues	
11	SOP & PTW	Equipment Issues	Safety Barrier/isolation issues	
12	SOP & PTW	Equipment Issues		
13	SOP & PTW	Safety Barrier/isolation issues		

5. Conclusion

Thus, this framework allows us to obtain chain of events from text using a dictionary of anchor words, applying semi-supervised topic model that identifies faults in form of topics with maximal information w.r.t the anchor words and finally developing the rules of association between the topics obtained. Since the topics represent the RCS failures, they represent the

various faults resulting in the incident. From Table 4, it can be seen that the major faults that are root cause of various incidents in the petrochemical plant are procedures, staff competency, equipment malfunction and safety barrier/energy isolation issues. Hence the management of the plant needs to focus on these faults to bring the risk in the plant under control.

6. Acknowledgement

The authors acknowledge the Centre of Excellence in Safety Engineering and Analytics (CoE-SEA) (www.iitkgp.ac.in/department/SE), IIT Kharagpur 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. The authors would like to thank the management of the plant for providing relevant data and for their support and cooperation during the study.

7. References

1. Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.
2. Ekramipooya, A., Boroushaki, M., & Rashtchian, D. (2023). Application of Natural Language Processing and Machine Learning in Prediction of Deviations in the HAZOP Study Worksheet: A Comparison of Classifiers. Process Safety and Environmental Protection.
3. Ertek, G., & Kailas, L. (2021). Analyzing a decade of wind turbine accident news with topic modeling. Sustainability, 13(22), 12757.
4. Gallagher, R. J., Reing, K., Kale, D., & Ver Steeg, G. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. Transactions of the Association for Computational Linguistics, 5, 529-542.
5. Heinrich, H. W. (1941). Industrial Accident Prevention. A Scientific Approach. *Industrial Accident Prevention. A Scientific Approach.*, (Second Edition).
6. Khan, F. I., & Abbasi, S. A. (1998). Techniques and methodologies for risk analysis in chemical process industries. *Journal of loss Prevention in the Process Industries*, *11*(4), 261-277.
7. Khan, F., Rathnayaka, S., & Ahmed, S. (2015). Methods and models in process safety and risk management: Past, present and future. *Process safety and environmental protection*, *98*, 116-147.
8. Meng, X., Zhu, J., Fu, J., Li, T., & Chen, G. (2021). An accident causation network for quantitative risk assessment of deepwater drilling. *Process Safety and Environmental Protection*, *148*, 1179-1190.
9. Qiu, Z., Liu, Q., Li, X., Zhang, J., & Zhang, Y. (2021). Construction and analysis of a coal mine accident causation network based on text mining. *Process safety and environmental protection*, *153*, 320-328
10. Roque, C., Cardoso, J. L., Connell, T., Schermers, G., & Weber, R. (2019). Topic analysis of road safety inspections using latent Dirichlet allocation: A case study of roadside safety in Irish main roads. *Accident Analysis & Prevention*, *131*, 336-349.
11. Robinson, S. D. (2019). Temporal topic modeling applied to aviation safety reports: A subject matter expert review. *Safety science*, *116*, 275-286.
12. Reason, J. (1990). *Human error*. Cambridge university press.
13. Sahoo, S., Maiti, J., & Tewari, V. K. (2024). A framework to model contractors' hazard and risk exposure at process plants using unsupervised text mining. *Process Safety and Environmental Protection*, *183*, 24-45
14. Sahoo, S., & Maiti, J. (2022, October). Identification of Accident Path Elements using Supervised Learning: Observation of Diminishing Marginal Accuracy while using Cosine Similarity. In *2022 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 135-139). IEEE.
15. Singh, K., Maiti, J., & Dhalmahapatra, K. (2019). Chain of events model for safety management: data analytics approach. *Safety science*, *118*, 568-582.
16. Singh, K., & Maiti, J. (2020). A novel data mining approach for analysis of accident paths and performance assessment of risk control systems. *Reliability Engineering & System Safety*, 202, 107041
17. Triplett, T. L., Zhou, Y., & Mannan, M. S. (2004). Application of chain of events analysis to process safety management. *Process safety progress*, *23*(2), 132-135.