Reverse HAZOP: Enhancing Safety Improvements through Natural Language Processing and Text Mining

Seyed Mojtaba Hoseyni a, Weixiang Han b, Joan Cordiner a\*

a University of Sheffield, Chemical and Biological Engineering, Sheffield S1 3JD, UK

b University of Sheffield, Computer Science, Sheffield S1 4DP, UK

\* [j.cordiner@sheffield.ac.uk](mailto:j.cordiner@sheffield.ac.uk)

Abstract

This research presents an intelligent method that uses text mining to enhance the effectiveness of Hazard and Operability Analysis (HAZOP). Despite HAZOP’s effectiveness in identifying hazards, its findings are often overlooked and conducting new analysis is time-consuming. Text mining, part of Natural Language Processing (NLP), can analyse safety-related data to identify risks and factors contributing to safety incidents, enabling proactive safety measures. The study introduces a HAZOP analysis data table structure to develop safety inspection checklists and proposes a reverse HAZOP analysis method. Case studies are used to demonstrate the practical application and benefits of NLP and text mining in safety improvements and reverse HAZOP. It highlights the benefits of utilising these technologies in terms of enhanced safety measures, proactive risk management, and improved decision-making.

**Keywords**: HAZOP, Text mining, Process safety, Risk analysis, Natural Language Processing (NLP)

* 1. Introduction

HAZOP, having a rich history in process safety since the 1960s, is a reputed method of hazard identification and risk analysis. Safety and risk-prevention is paramount for industrial enterprises. Thus, identification and reflection of accidental faults is essential for constructing preventative control systems which is achieved by conducting HAZOP (Martínez, et al., 2019). Despite its esteem, HAZOPs findings and recommendations from published analysis reports are often overlooked, and therefore not applied. Moreover, the current process of conducting a new HAZOP analysis is often deemed as time-consuming and subjective (Pasman & Rogers, 2016).

On the contrary, NLP employs text mining algorithms to promise organisations incisive acquirement of safety related data from various sources- namely incident, safety and HAZOP reports and maintenance logs. Detection of hidden patterns is optimised, to highlight the contributing factors of safety incidents. Ultimately, it is used as a tool for mitigating proactive developments of robust safety guidelines- bringing forth training needs, and the implementation of preventive maintenance strategies (Robinson, et al., 2021).

This study develops a framework to execute the reverse HAZOP analysis method to validate existing studies. In this paper, real-world examples are used to illustrate the practical application of NLP and text mining techniques in the domain of safety improvements and reverse HAZOP.

The novelty of the research lies in introducing an intelligent approach that utilizes text mining, particularly NLP, to enhance HAZOP. Unique contributions include the development of safety inspection checklists based on a structured HAZOP analysis data table and the proposal of a reverse HAZOP analysis method. These innovations enable the identification of risks and contributing factors to safety incidents, facilitating proactive safety measures. The integration of NLP and text mining enhances decision-making and proactive risk management in industrial settings, marking a distinctive advancement in safety improvement methodologies.

* 1. Methodology

A novel approach to reverse HAZOP analysis, utilizing data mining techniques, is proposed in this study to uncover intricate relationships within comprehensive HAZOP data sets. These datasets encompass various elements such as process parameters, causes, consequences, guide words, probabilities, severities, and risk levels associated with potential hazards and accident scenarios in the HAZOP report. By computing word frequencies for process parameters, guide words, causes, and consequences, significant attention factors are identified based on high word frequencies. Furthermore, an information inspection table is devised to establish the inspection order. The methodology is structured into three integral components: (1) implementation of text data mining based on the HAZOP report, and (2) word frequency statistical analysis and (3) correlation analysis.

* + 1. **Data mining on the HAZOP text**

Because the analysis results of HAZOP are usually recorded in the form of report text, text data mining can extract the key information from its unstructured, qualitative data (Dang & Ahmad, 2014) - generating it as parameters, deviation safeguards etc. Text data mining is used to prompt the construction of the HAZOP analysis table with the following steps:

* + - 1. **Data Preprocessing**

Data preprocessing is a crucial step in the data mining and data analysis process of the HAZOP analysis table execution. Raw text data from the HAZOP report data needs to be transformed into a digestible, uniform design. Unstructured data, in the form of text, must first be cleaned and formatted before any conclusive analysis and modelling can be made. Such process involves the manipulation, or dropping of data before it is used to ensure or enhance performance (Jansen, et al., 2023).

* + - 1. **Text Clustering Analysis – Latent Dirichlet Allocation (LDA) Model**

Latent Dirichlet Allocation (LDA) is a widely-used form of statistical topic modelling- treating documents as a probabilistic distribution sets of words or topics (Jelodar, et al., 2019). Classification is used to succinctly categorise the text into a document, and the words per topic (Blei, et al., 2003). Documents are represented as a mixture of topics, and a topic as a bunch of words. Those topics reside within a hidden/latent layer- hence the name.

Despite its nature, LDA is not a clustering algorithm. Whilst both methods group documents, its process differ. The LDA topic model is based on the intuition that words that belong to a topic tend to appear together in documents (Blei, 2012).

* + 1. **Word Frequency Statistic**

Following the investigation and analysis of the HAZOP nodes, word frequency analysis needs to be explored. Such statistical analysis is commonly used to identify the most frequently used words in qualitative text reports. It is seen throughout various applications- namely NLP, but also search engine optimisation, and content analysis. As such, it can help to highlight key themes and trends. This information can be used to improve the readability and clarity of the report, and therefore identify areas that may require further investigation or clarification (Feng, et al., 2021).

After counting how many times each distinct word appears in a lengthy text, the words are arranged in decreasing order of frequency. Every word in the sample has its rank assigned in that order, and the product of the rank and the frequency is determined to be roughly equal for each word.

* + 1. **Correlation Analysis**

Correlation analysis is another statistical method. In the context of word frequency statistics, correlation analysis can be used to identify and evaluate the relationship/strength between the frequency of different words (variables) in a text.

If a strong positive correlation is determined, the frequency of the word is similar in both texts. Conversely, if there is a strong negative correlation, the frequency of the word is different between the texts (Ricketts, et al., 2022).

* 1. Results and Discussion

We chose a HAZOP report of a gas turbine unit as a case study. Data tables are built to the nodes of the case study. It briefs the key features relevant to standard HAZOP measures- parameters, guidewords, deviation, causes, consequences, severity, safeguards, likelihood, risk, ref, recommendations. 6 out of the total 46 nodes were selected for this study. The occurrence frequency of each node classification can reveal potential risks associated with specific nodes for plants. These risks can be quantified based on the data table from the HAZOP analysis. As shown in Figure 1, Node 02B and 08 in general, have occurrence time more than 100 times.

A graph of blue bars

Description automatically generated with medium confidence

**Figure 1**. Statistical results of equipment classification

To find the most relevant process parameters and guide words for each node in the HAZOP data table, frequency of each words can be counted and subsequently the words can be sorted by their frequency and use to form deviations. This way, missing any important information can be avoided when analysing the nodes. Figures 2 and 3 show the word frequency statistics of the process parameters and guide words for the nodes respectively.

A graph of a number of blue rectangular bars

Description automatically generated with medium confidenceA graph of blue rectangular bars

Description automatically generated with medium confidence

**Figure 2.** Process parameters’ word frequency

A graph of different colored columns

Description automatically generated

**Figure 3.** guide words’ word frequency

The frequency with which each word occurred in the cause and consequence of the node was counted using the word frequency statistics technique. It is suggested that in order to improve the visibility of the analysis results in this study, at least 20 terms with the highest frequency be chosen as keywords. Table 1 presents the terms in a frequency-descending order to facilitate a better understanding of the causes and effects.

**Table 1.** Cause and consequence of ‘Node 02’

A screenshot of a computer

Description automatically generated

Typically, the LDA model is evaluated using perplexity as an index. Figure 4's curve indicates that the number of topics at 7 is at the inflection point; as a result, 7 is the number of clustering subjects that are selected.

A graph with a line

Description automatically generated

**Figure 4.** Perplexity Curve

Using clustering analysis, the LDA model is utilised to count the percentage of the seven topics and cluster all of Node 02's causes and effects. Word frequency probabilities are computed, and the top 10 keywords for each topic are used to represent it and arrange the results. Table 2 shows consequence topics for Node 2 with the word frequency probabilities associated with the specific keywords of the consequence topic. Topic names can be defined by looking up each topic's keyword usage.

**Table 2.** The 'Node 02' consequence topics' clustering analysis

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Consequence Topic | Keywords | Probability |
| 0 | Piping Failures | Possible, Pressure, May, Back, Fatalities, Less, Design, Piping, Downstream, Process | 0.46 |
| 1 | Flow related anomalies | Flow, Leading, Gas, Loss, Increasing, Scenario, Hazardous, Partial, Production, Wet | 0.65 |
| 2 | Compressor failures | Pressure, Compressor, Suction, Possible, Damage, Low, Surge, Leading, Flow, Explosion | 0.16 |
| 3 | Equipment damage | Compressor, Suction, Flow, Leading, Pressure, Possible, Damage, Surge, Low, Explosion | 0.80 |
| 4 | Pressure related anomalies | Pressure, Possible, Compressor, Damage, Surge, Suction, Low, Flow, Explosion, Leading | 0.16 |
| 5 | Medium failure with Interrupted operation | Compressor, Suction, Pressure, Low, Surge, Damage, Possible, Leading, Flow, Lead | 0.92 |
| 6 | Severe failure | Pressure, Possible, Explosion, Damage, Surge, Compressor, Potential, Fatalities, May, Fire | 0.48 |
| 7 | Abnormal level in suction drum | Pressure, Possible, High, May, Fire, Rupture, Exceed, Upstream, Equipment, Leak | 0.62 |

A contingency table is used to determine whether two aspects of the same analysis data are correlated. From Figure 5, process parameters and guidewords are used. The results show that ‘No/ Less’ happened the most in ‘Flow’, following ‘Less’ in ‘Level’.

A screenshot of a computer

Description automatically generated

**Figure 5.** Process parameters and guidewords of ‘Node 02’

* 1. Conclusions

This study introduces an intelligent approach leveraging text mining techniques to augment the efficiency of Hazard and Operability (HAZOP) analysis. The HAZOP analysis data table served as the foundation for implementing both a word frequency statistical algorithm and Latent Dirichlet Allocation (LDA) clustering models. The interrelationships among various HAZOP elements are illustrated, offering valuable insights for safety management personnel. This information can be instrumental in the analysis, prevention, and control of accidents within chemical plants. The findings have substantial implications for enhancing the safety and reliability of production processes in industrial plants. In evaluating the practical application of the proposed approach, it is essential to acknowledge potential limitations and challenges. The generalizability of the findings may be constrained by the specific context and dataset used in this study. Further research and validation across diverse industrial settings are warranted to establish the broader applicability and robustness of the intelligent approach. Addressing these concerns will contribute to a more comprehensive understanding of the method's effectiveness and facilitate its successful implementation across various HAZOP analysis scenarios.

References

Blei, D., 2012. Probabilistic topic models. *Communications of the ACM,* 55(4), pp. 77-84.

Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003. Latent dirichlet allocation. *Journal of machine Learning research,* Volume 3, pp. 993-1022.

Dang, S. & Ahmad, P. H., 2014. Text mining: Techniques and its application. *International Journal of Engineering & Technology Innovations ,* 1(4), pp. 22-25.

Feng, X. et al., 2021. Application of natural language processing in HAZOP reports. *Process Safety and Environmental Protection ,* Volume 155 , pp. 41-48..

Jansen, B. J. et al., 2023. Data Preprocessing. In: *Understanding Audiences, Customers, and Users via Analytics: An Introduction to the Employment of Web, Social, and Other Types of Digital People Data.* s.l.:Springer, Cham.

Jelodar, H. et al., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications,* Volume 78, pp. 15169-15211.

Martínez, B., Rodríguez, M. & Díaz, I., 2019. Automating HAZOP studies using D-higraphs. *Computer Aided Chemical Engineering ,* Volume 46, pp. 553-558.

Pasman, H. J. & Rogers, W. J., 2016. How can we improve HAZOP, our old work horse, and do more with its results? An overview of recent developments. *Chemical Engineering,* Volume 48, pp. 829-834.

Ricketts, J., Pelham, J., Barry, D. & Guo, W., 2022. *An NLP framework for extracting causes, consequences, and hazards from occurrence reports to validate a HAZOP study.* s.l., s.n.

Robinson, C., Brown, S. & Cordiner, J., 2021. Towards Automated HAZOPs. *Computer Aided Chemical Engineering,* Volume 50, pp. 505-510.