[1] Yunan Zhang*

[2] Xiaoyu Wang

A Decision Tree Algorithm in University Laboratory Hazardous Chemicals Management and Alternative Technology Research

Abstract: - Laboratory-based research activities include inherent hazards associated with managing hazardous chemicals and processes, requiring comprehensive risk assessment protocols to prevent accidents and injuries. In this study, we investigate the effectiveness of decision tree (DT) algorithms in detecting possible hazards in laboratory work using previous accident documents. The dataset consists of accidents and near-miss incidents reported from various university laboratory operations. Each report describes the nature of laboratory operations and related hazards. In our proposed model, we employ a Word2Vec-based DT algorithm, where Word2Vec transforms words into semantic vectors. The DT algorithm then utilizes these vectors to estimate hazard probabilities by recursively partitioning the data according to their characteristics for predicting the risks associated with university laboratory work. The proposed detection model has been implemented in a Python program. In the results assessment phase, we evaluate our proposed model's effectiveness in forecasting various chemical hazardous situations using numerous evaluation metrics such as recall, f1 score, precision and accuracy. We also carried out a comparison analysis with other traditional approaches. Our experimental findings demonstrate the reliability of the recommended framework.

Keywords: Laboratory Hazardous Chemicals, Decision Tree (DT), Word2Vec-based DT Algorithm, University Laboratory Works.

1. Introduction

University-conducted innovative experiments can involve the use of dangerous instruments or laboratory procedures [1]. Additionally, they could be involved in risky tasks including handling pyrophoric materials, inactivating infectious pathogens, moving large gas cylinders, and completing metalwork with machine tools which are highly probable to end in accidents and near-miss occurrences [2].

The complexity of laboratory safety management grows when there is a chance of fire, explosion and other issues. It is risky to undertake laboratory research given the rise in accident frequency [3]. Many different types of laboratories, including biological, chemical, electrical, mechanical and environmental labs, are frequently found in one university. Every kind of laboratory is made up of several rooms with various purposes. Additionally, every area has a variety of tools and apparatus. This implies that running the university laboratory presents significant difficulties [4]. In actuality, issues with management work account for the majority of laboratory accidents. Daily management tasks depend on the equipment's routine inspection and maintenance to guarantee the safety of the laboratory. Manual statistics are typically used in the administration of the laboratory apparatus [5]. The task of equipment management is difficult, time-consuming, and costly since there are so many different kinds of equipment.

Faculty personnel, researchers, graduate students, or students could be killed in an incident that occurs in a university laboratory while conducting a chemical experiment [6]. The number of laboratory accidents at universities worldwide is unknown, but after they happened, similar incidents occurred again at other colleges [7]. In these instances, there wasn't a significant paradigm shift or change in the way laboratories are safe. Furthermore, the colleges' efforts to work together to stop similar incidents have not advanced [8].

The unusual and extensive attention from industry and mainstream media, as well as from academic institutions around the country, complicated the campus's response efforts even further [9]. The incident had a major impact on everyone in school, especially those who work in laboratory safety and research, the office of Sustainability, Health & Safety (EH&S) is responsible for safeguarding the health and safety of all individuals on campus. [10]. We utilize a Word2Vec-based DT method in this work, which converts words into semantic vectors. Next, by recursively dividing the data based on these vectors, the DT method makes use to estimate hazard probability to anticipate the dangers related to university laboratory work.

The following is the order of this paper: Related works are included in section 2, and techniques are covered in section 3. Section 4 presents the experiment's results. Section 5 concludes with a summary of the study and suggestions for more research.

2. Related works

The inherent hazards assessment and categorization (IHAC) approach was developed in study [11] for use in university chemistry labs as a result of this investigation. They conducted a quantitative assessment of the

[1,2] National Center for Materials Service Safety, University of Science and Technology Beijing, Beijign 100083, Beijing, China. *E-mail: zhangyn85@163.com

inherent risks in chemical laboratory materials, apparatus and procedures. Concurrently, dividing labs into many levels can lead to more focused safety management and accident avoidance.

Seven substances' chemical concentrations were used in research [12], to examine the fluctuations in potential of Hydrogen (pH) and Electrical conductivity (EC). The seven compounds were known to produce chemical spills regularly in South Korea and that were classified as accident-preparedness substances. Furthermore, they compared the changes in pH, EC, and statistics during the dilution procedure to determine the probability of recognizing unknown chemicals.

The process of developing Standard Operating Procedures (SOPs) in research [13] afforded the chance to ascertain the necessary conditions for reaction setup, recognize possible dangers, define the appropriate handling of undesired materials, and conduct a comprehensive risk assessment. Here, they offered recommendations for SOPs that have to be created for university research facilities as well as an example of an SOP for the Grignard reaction.

A technique for the prospective evaluation of chemicals using sorting-based multi-parameter and multi-criteria decision-making (MCDM) hazards to worker safety in the university lab was established in research [14]. It was advised that certain control measures should be implemented to lower the laboratory's safety risk. The technique was meant to become a main source of information for university danger analysts and adjust to the risk assessment of university laboratories.

The fuzzy Bayesian network (BN) method combined with the human factors assessment and categorization system for university labs (HFACS-UL) was suggested in study [15], to evaluate the risky conduct in university laboratories. The primary risk factors were determined by applying the model to an inference analysis. To identify further preventative and control methods, meta-networks and important agents were also investigated.

To discover the implementation of certain semi-quantitative methods was assessed in research [16], to determine potential bias or variances caused by utilizing different ways to the same tasks for chemical risk assessment. They could overcome the discrepancies observed in the risk assessment by using two or more distinct semi-quantitative instruments for every working activity they need to evaluate. The tactic could allow workers' contact with chemicals to be reduced.

After an analysis of statistical information to provide a broad explanation of the traits of greater and more frequent accidents, research [17] estimated and evaluated the total risk of the hazardous chemical sector using the entropy weight technique. It examined how safety laws have evolved in China's hazardous chemical business and how the sector was expected to grow moving forward.

The analytical tool based on the software, hardware, environment and liveware (SHEL) paradigm was utilized in study [18], to examine reports from accident investigations about explosions at two universities' chemical labs. Global university communities must collaborate to develop methods for research and analysis, instructions for writing accident reports and an information-sharing platform that would enable them to take advantage of the knowledge gathered from a range of incidents.

Hazardous chemical control was a vital component of campus laboratory safety management, as demonstrated through study [19] and subsequent investigation of remedies and countermeasures. Realizing the intrinsic protection of the university laboratory, the safety management method was completed with the construction of the basis for safety management for its whole life of hazardous substances.

An explosive accident at a university laboratory was thoroughly investigated in study [20], to determine the primary reasons and enhance safety management. The findings suggested that the experimenters' lack of caution and poor safety knowledge were the primary causes of the accident. To successfully prevent these kinds of tragedies, experimenters and related technical managers need to receive more safety training. It would help to foster a positive safety culture inside the institution.

It focused on methods for deciphering and discovering the possible reasons behind the actions of the people implicated in mishaps in chemical laboratories in study [21]. Reflections could be beneficial for a variety of stakeholders, including administrators, graduates, suppliers and producers of chemicals and lab equipment, managers, universities and colleges making investments in new or renovated chemical laboratories, environment, safety, and health (ESH) professionals.

There were many employment contexts, where people were exposed to chemicals in study [22], but the investigation and healthcare facilities have not received enough attention. It examined how research laboratory staff were exposed to hazardous chemicals at work, evaluated their knowledge and attitudes about chemical hazards, examined whether they followed the rules for handling chemicals safely, and examined the impact of various factors on the important outcomes.

3. Methodology

The collection includes recorded near-miss and accident events from a variety of academic laboratory operations. Using a Word2Vec-based DT algorithm which converts words into semantic vectors, we implement this approach in our suggested model. To anticipate the dangers connected with university laboratory work, the DT method uses these vectors to recursively segment the data based on their features, estimating hazard probabilities. The general flow is depicted in Figure 1.

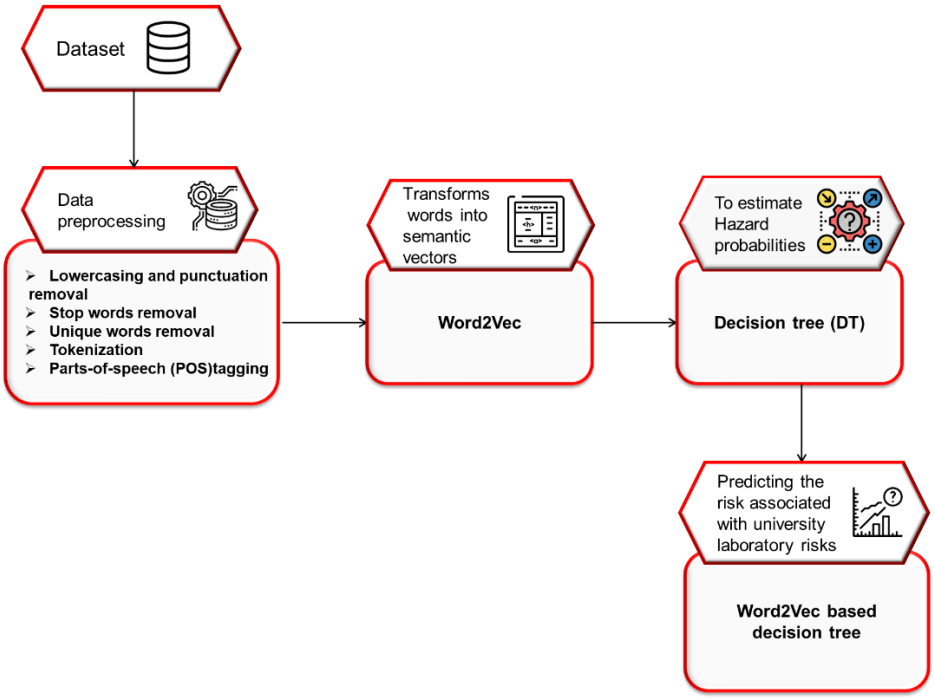

Figure 1: Overall flow

3.1 Dataset

A synthetic dataset consisting of event reports from university laboratory activities was created. Sample analysis, equipment maintenance and chemical synthesis are few of the laboratory tasks included in these reports. All of the reports consist of records on the dangers that would rise from those activities, inclusive of sample analysis, chemical synthesis and equipment maintenance. The dataset accommodates 1133 reports which have been randomly categorized as either close to pass-over occurrences or accidents to copy various levels of severity. This dataset is the foundation of research on DT algorithms identify risks in lab operations.

3.2 Text preprocessing

Stop Words Removal: Stop words are often used words like "the," "and," "is," "for," and "of" in statements that lack distinction and significance. To lower the noise in the model, every English stop word in the corpus was eliminated for this investigation.

Lowercasing and Punctuation Removal: The entire text was transformed to lowercase using the lowercasing technique to guarantee that terms with similar meanings, such as "Construction" and "construction," are regarded as a single phrase. As they had no bearing on the text's categorization, the punctuation marks were similarly eliminated.

Unique Words Removal: Preprocessing also eliminated unique terms or extremely low-frequency words that appeared once across the corpus since they lacked discriminative ability.

Tokenization: Tokenization divides the lengthy, full text into discrete pieces, or tokens, which could consist of a single word, a period, an integer, or a white space.

Parts-of-speech (POS) tagging: The terms developing and stairs were labelled as (NN) singular noun, selected and brick-built as (JJ) adjectives, had and has as (VB) verbs, and so on. The tokens were allocated POS tags that corresponded to the useful and lexical classifications of the terms or tokens.

3.3 Word2Vec

Popular word embedding method Word2Vec could be used to manage dangerous compounds in the lab by encoding chemical names and qualities into high-dimensional vectors. This makes it easier to do similarity and group chemicals according to their attributes. The continuous distributed display of words is computed using the Word2Vec model. This model offers an effective way to compute vector representations of words using the Continuous Bag of Words Model (CBOW) and Skip-gram designs, which are both basic neural network models with a single hidden layer. The Word2Vec model accepts a text corpus as input and outputs word vectors. Using backpropagation and stochastic descent of gradients, this approach first creates a vocabulary derived from the words entered. After that, the word vectors are learned. The CBOW architecture predicts a word from future and previous contexts by utilizing a log-linear classifier learned through the negative collection and averaging contextual word vectors. With a given word, the Skip-gram architecture predicts surrounding words. Training examples are created by eliminating a predetermined number of contextual phrases since the context is unbounded, such as $x_j - 3, x_j - 4, x_j + 3, x_j + 4$, thus the term "skip-gram." Figure 2 shows the CBOW and skip-gram structure.

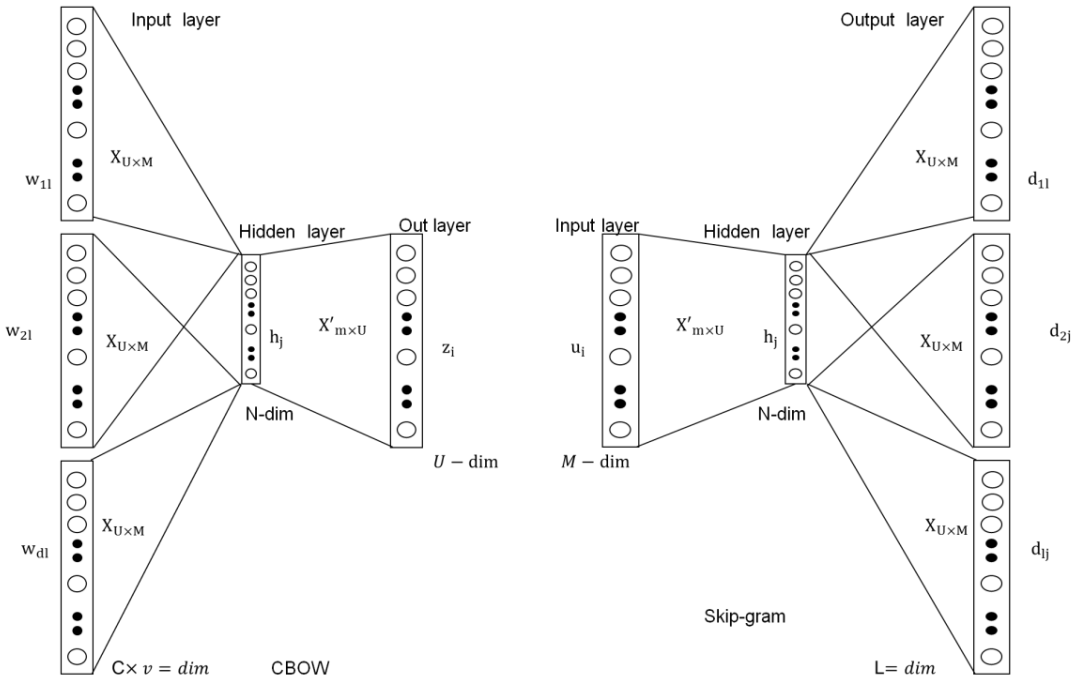

Figure 2: CBOW and skip-gram

Instead of creating a new model from scratch, the pre-trained model is utilized to address a comparable scenario or issue. Pre-trained models are developed and made available to the public for use in research. Pre-trained Word2Vec and Pre-trained GloVe are the two pre-trained models that are available for word embedding.

3.4 Decision tree

The DT method uses these vectors to recursively split the data based on their features to determine hazard probability and anticipate the dangers related to the university chemical lab. The DT has branches containing qualities that determine the outcome, or the objective function and organized in a sequential hierarchical structure. Nodes: arbitrary vertices where the potential course of events is ascertained, the outcome of Leaf (leaf) nodes with intends and values are used to depict the process of choosing a certain attribute value and merging several objects. Depending on the type of predicted indicator, decision trees could be divided into two categories: regression trees and classification trees. Trees of classification are useful for study on certain qualities, such as assigning items to a previously established class hence using them is advised when creating a prediction system. Data is categorized using decision trees, which split data into groups and provide a hierarchy of "if... then..." operators.

To separate the nodes into informative functions, create an objective function. Every division in which we optimize the increase is:

$$JH(C_o, e) = J(C_o) - \sum_{i=1}^{n} \frac{M_i}{M_o} J(C_i) \quad (1)$$

Where e is the property that is used to conduct the splitting; Parents C_o and C_i is the $i-th$ child nodes, while J is a heterogeneity measure. M_i is the quantity of specimens contained in the $i-th$ child node, M_o is the overall number of values in the parent node.

We use binary decision trees for simplicity and to shrink the multidimensional search space. The child nodes $C_{left}$and C_{right} in our scenario are:

$$JH(C_o, e) = J(C_o) - \sum_{j=1}^{d} \frac{M_{left}}{M_o} J\left(C_{left}\right) - \frac{M_{right}}{M_o} J\left(C_{right}\right) \quad (2)$$

Where e is the property that is used to conduct the splitting; the parent and $i-th$ child node databases are denoted by $J(C_o)$. J is a heterogeneity metric; The total number of samples in the parent node is represented by C_o, the number of samples in the child nodes that are in the $i-th$ child node is represented by M_o, the child nodes' numbers of patterns are represented by M_{left} and M_{right}.

Determination of entropy for all classes $o(j/s) \neq 0^2$ that is not empty.

$$J_G(s) = -\sum_{j=1}^{d} o(j/s) log_2 o(j/s) \quad (3)$$

Where $o(j/s)$ is the percentage of samples including a single node s and the class.

Therefore, if every sample in a node is a member of the same class, the entropy is zero, and if the distribution of classes is uniform, the entropy is at its maximum.

A criterion reduces the possibility of misdiagnosis is the Gini statistic for heterogeneity:

$$J_H(s) = -\sum_{j=1}^{d} o(j/s)(1 - o(j/s)) = 1 - \sum_{j=1}^{d} o(j/s)^2 \quad (4)$$

Where $K_H(s)$ is the Gini measure of heterogeneity and $o(j/s)$ is the proportion of samples that fall under a class and a single node.

Classification error is an additional metric for heterogeneity.

$$J_\varepsilon(s) = 1 - max\{o(j/s)\} \quad (5)$$

Where s is the single node and $o(j/s)$ is the proportion of samples that correspond to a class, $J_\varepsilon(s)$ is the classifier error.

Although this criterion is less susceptible to changes in the capacity of the classes at the nodes, it is appropriate for trimming trees but not for growing trees.

3.5 Word2Vec-based DT

An innovative technique for coping with dangerous substances in academic labs is the Word2Vec-based DT set of rules. This technique turns chemical descriptions and protection information into excessive-dimensional vectors that seize the complicated interactions between terms by using Natural language processing (NLP) and ML. The gadget can realise links among materials, risks, and safety regulations for the reason of those vectors encode linguistic commonalities. The programs benefit in categorizing compounds in step with their traits and associated dangers by means of training a DT model on these vector representations. The risky chemical machine gives more precise control and is streamlined by using this automatic class process, which reduces the need for human inspection and expertise. Furthermore, the machine gives extra particular hazard critiques and customized safety advice by using the semantic context under consideration. All things taken into consideration, there may be a great deal of capacity for elevating protection standards and decreasing dangers in university laboratories through the use of this novel technique.

4. Experimental results

The recommended method is implemented on a Windows 10 laptop with an Intel i7 core CPU and 8GB of RAM using Python 3.10.1. Utilize frameworks such as Scikit-Learn or Tensor Flow/Keras to train our suggested model on the training set. The methods are Word2Vec, DT, and Word2Vec-based DT. The performance measures include f1-score, accuracy, precision, and recall.

Accuracy in the context of managing hazardous chemicals in university laboratories refers to a detection model's capacity to identify both safe operations and hazardous situations. It is a metric of the model performs overall in terms of accurately recognizing non-hazardous circumstances and forecasting real hazardous scenarios.

Figure 3 and Table 1 display the accuracy performance. The methods are Word2Vec-based DT, Word2Vec, and DT which achieve an accuracy of 92.3%, 85%, and 88%. Consequently, the Word2Vec-based DT is more accurate than the other methods used for Hazardous chemical handling at university laboratories.

Table 1: Values for precision, accuracy, recall, and F1 score

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Word2Vec	85	88	84.2	82
DT	88	83.4	86.7	86
Word2Vec based DT [Proposed]	92.3	93.1	91.4	93.5

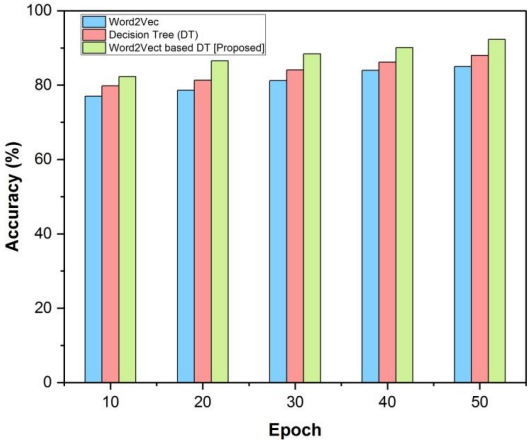

Figure 3: Accuracy performance

Precision in university hazardous chemical management refers to the system's reliability in accurately recognizing hazardous chemical-related situations. Specially, it calculates the percentage of actual positive danger detections among all the systems detections.

The precision performance is shown in Figure 4 and Table 1. The approaches, which provide values of 93.1%, 88%, and 83.4%, are Word2Vec based DT, Word2Vec, and DT. As a result, while handling hazardous chemicals in university laboratories, the Word2Vec-based DT is more precise than the other techniques.

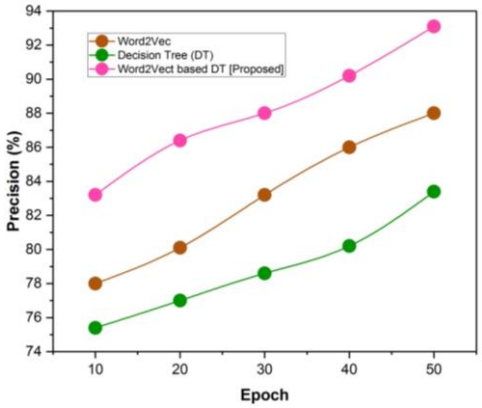

Figure 4: Precision performance

When discussing university laboratory hazardous chemicals management, recall refers to the statistic used to assess a detection model identifies real hazardous situations. It can be defined as the proportion of properly classified hazardous occurrences on all actual hazardous incidents.

Recall performance is displayed in Table 1 and Figure 5. The methods Word2Vec-based DT, Word2Vec, and DT offer values of 91.4%, 84.2%, and 86.7%. Therefore, the Word2Vec-based DT is more reliable than the other methods for handling hazardous substances in university laboratories.

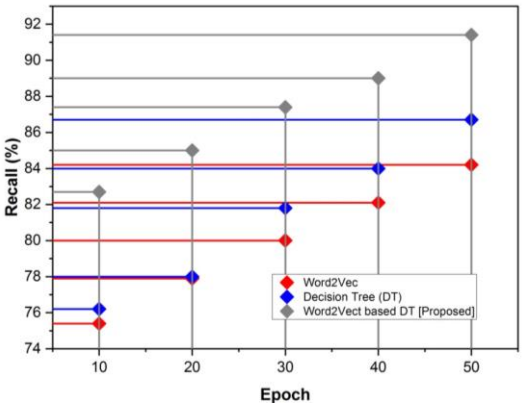

Figure 5: Recall performance

An important indicator for assessing hazard detection models' performance in university laboratory chemical control is the F1 score. It is the precision and recall harmonic mean that balances false positives and false negatives.

The F1 score results are shown in Table 1 and Figure 6. Values of 93.5 %, 82%, and 86% are attained using Word2Vec based DT, Word2vec, and DT techniques. Therefore, compared to other techniques used for managing hazardous chemicals in university laboratories, the Word2Vec-based DT is better.

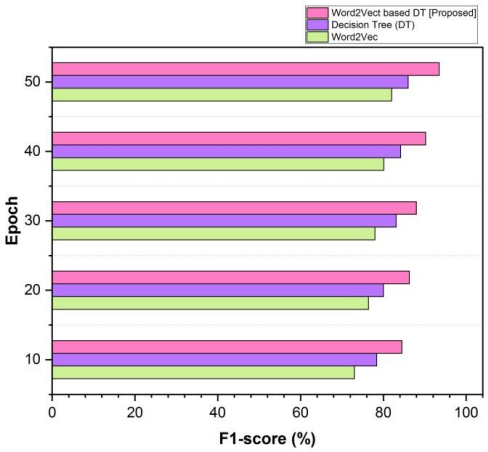

Figure 6: F1 score performance

5. Conclusion

In this work, we used accident reports from before to examine DT algorithms discover potential dangers in lab paintings. The collection is made from near miss and accident reports from several instructional laboratory operations. In this study, we used accident reports from earlier to have a look at how nicely DT algorithms identify capability hazards in lab work. The collection was made up of near-miss and accident reports from a variety of academic laboratory operations. The nature of laboratory activities and associated hazards are included in each report. We used a Word2Vec-based DT technique in our suggested model, which converts words into semantic vectors. The DT method uses these vectors to recursively split the data based on their features to estimate hazard probabilities and anticipate the dangers related to university laboratory work. When compared to the existing method, the proposed method achieves accuracy (92.3%), precision (93.1%), F1 score (93.5%), and recall (91.4%), respectively. Although the Decision Tree Algorithm simplifies the handling of hazardous chemical compounds in college laboratories, other developments in the vicinity want to consist of predictive analytics and non-stop surveillance for proactive risk discount and efficient use of sources.

Reference

[1] Ostad-Ali-Askari, K., 2022. Management of risks substances and sustainable development. Applied Water Science, 12(4), p.65.

[2] Gopalaswami, N. and Han, Z., 2020. Analysis of laboratory incident database. Journal of Loss Prevention in the Process Industries, 64, p.104027.

[3] Okebukola, P.A., Oladejo, A., Onowugbeda, F., Awaah, F., Ademola, I., Odekeye, T., Adewusi, M., Gbeleyi, O., Agbanimu, D., Peter, E. and Ebisin, A., 2020. Investigating chemical safety awareness and practices in Nigerian Schools. Journal of Chemical Education, 98(1), pp.105-112.

[4] Nasrallah, I.M., El Kak, A.K., Ismaiil, L.A., Nasr, R.R. and Bawab, W.T., 2022. Prevalence of accident occurrence among scientific laboratory workers of the public university in Lebanon and the impact of safety measures. Safety and health at work, 13(2), pp.155-162.

[5] Zhu, C., Tang, S., Li, Z. and Fang, X., 2020. Dynamic study of critical factors of explosion accident in laboratory based on FTA. Safety Science, 130, p.104877.

[6] Bai, M., Liu, Y., Qi, M., Roy, N., Shu, C.M., Khan, F. and Zhao, D., 2022. Current status, challenges, and future directions of university laboratory safety in China. Journal of Loss Prevention in the Process Industries, 74, p.104671.

[7] Li, X., Zhang, L., Zhang, R., Yang, M. and Li, H., 2021. A semi-quantitative methodology for risk assessment of university chemical laboratory. Journal of Loss Prevention in the Process Industries, 72, p.104553.

[8] Galasso, A., Luo, H. and Zhu, B., 2023. Laboratory safety and research productivity. Research Policy, 52(8), p.104827.

[9] Ezenwa, S., Talpade, A.D., Ghanekar, P., Joshi, R., Devaraj, J., Ribeiro, F.H. and Mentzer, R., 2022. Toward improved safety culture in academic and industrial chemical laboratories: an assessment and recommendation of best practices. ACS Chemical Health & Safety, 29(2), pp.202-213.

[10] Schröder, I., Czornyj, E., Blayney, M.B., Wayne, N.L. and Merlic, C.A., 2020. Proceedings of the 2018 laboratory safety workshop: hazard and risk management in the laboratory. ACS Chemical Health & Safety, 27(2), pp.96-104.

[11] Liu, S., Ju, S., Meng, Y., Liu, Q. and Zhao, D., 2023. Inherent Hazards Assessment and Classification Method for University Chemical Laboratories in China. ACS Chemical Health & Safety, 30(4), pp.156-164.

[12] Nam, S.H., Ku, T.G., Park, Y.L., Kwon, J.H., Huh, D.S. and Kim, Y.D., 2022. Experimental study on the detection of hazardous chemicals using alternative sensors in the water environment. Toxics, 10(5), p.200.

[13] Chandra, T., Zebrowski, J.P., McClain, R. and Lenertz, L.Y., 2020. Generating standard operating procedures for the manipulation of hazardous chemicals in academic laboratories. ACS Chemical Health & Safety, 28(1), pp.19-24.

[14] Gul, M., Yucesan, M. and Karacahan, M.K., 2023. A Multi-parameter Occupational Safety Risk Assessment Model for Chemicals in the University Laboratories by an MCDM Sorting Method. In Advances in Reliability, Failure and Risk Analysis (pp. 131-149). Singapore: Springer Nature Singapore.

[15] Li, Z., Wang, X., Gong, S., Sun, N. and Tong, R., 2022. Risk assessment of unsafe behavior in university laboratories using the HFACS-UL and a fuzzy Bayesian network. Journal of safety research, 82, pp.13-27.

[16] Mastrantonio, R., Scatigna, M., D'Abramo, M., Martinez, V., Paoletti, A. and Fabiani, L., 2020. Experimental application of semi-quantitative methods for the assessment of occupational exposure to hazardous chemicals in research laboratories. Risk Management and Healthcare Policy, pp.1929-1937.

[17] Yang, D., Zheng, Y., Peng, K., Pan, L., Zheng, J., Xie, B. and Wang, B., 2022. Characteristics and statistical analysis of large and above hazardous chemical accidents in China from 2000 to 2020. International journal of environmental research and public health, 19(23), p.15603.

[18] Fukuoka, K. and Furusho, M., 2022. A new approach for explosion accident prevention in chemical research laboratories at universities. Scientific reports, 12(1), p.3185.
[19] Yang, J., Xuan, S., Hu, Y., Liu, X., Bian, M., Chen, L., Lv, S., Wang, P., Li, R., Zhang, J. and Shu, C.M., 2022. The framework of safety management in a university laboratory. Journal of Loss Prevention in the Process Industries, 80, p.104871.
[20] Chen, M., Wu, Y., Wang, K., Guo, H. and Ke, W., 2020. An explosion accident analysis of the laboratory in the university. Process Safety Progress, 39(4), p.e12150.
[21] McLeod, R.W., 2022. Approaches to Understanding Human Behavior When Investigating Incidents in Academic Chemical Laboratories. ACS Chemical Health & Safety, 29(3), pp.263-279.
[22] Papadopoli, R., Nobile, C.G.A., Trovato, A., Pileggi, C. and Pavia, M., 2020. Chemical risk and safety awareness, perception, and practices among research laboratories workers in Italy. Journal of occupational medicine and toxicology, 15, pp.1-11.