SPE-205877-MS

Cognitive HSE Risk Prediction and Notification Tool Based on Natural Language Processing

Tharunya Danabal, Neethi Sarah John, Abhijeet Pramod Ghawade, and Pranjal Padharinath Ahire, Schlumberger

This paper was prepared for presentation at the 2021 SPE Annual Technical Conference and Exhibition held in Dubai, UAE, 21 - 23 September 2021.

Abstract

The focus of this work is on developing a cognitive tool that predicts the most frequent HSE hazards with the highest potential severity levels. The tool identifies these risks using a natural language processing algorithm on HSE leading and lagging indicator reports submitted to an oilfield services company's global HSE reporting system. The purpose of the tool is to prioritize proactive actions and provide focus to raise workforce awareness.

A natural language processing algorithm was developed to identify priority HSE risks based on potential severity levels and frequency of occurrence. The algorithm uses vectorization, compression, and clustering methods to categorize the risks by potential severity and frequency using a formulated risk index methodology. In the pilot study, a user interface was developed to configure the frequency and the number of the prioritized HSE risks that are to be communicated from the tool to those employees who opted to receive the information in a given location.

From this pilot study using data reported in the company's online HSE reporting system, the algorithm successfully identified five priority HSE risks across different hazard categories based on the risk index. Using a high volume of reporting data, the risk index factored multiple coefficients such as severity levels, frequency and cluster tightness to prioritize the HSE risks. The observations at each stage of the developed algorithm are as follows:

- In the data cleaning stage, all stop words (such as a, and, the) were removed, followed by tokenization to divide text in the HSE reports into tokens and remove punctuation.
- In the vectorization stage, many vectors were formed using the Term Frequency - Inverse Document Frequency (TF-IDF) method.
- In the compression stage, an autoencoder removed the noise from the input data.
- In the agglomerative clustering stage, HSE reports with similar words were grouped into clusters and the number of clusters generated per category were in the range of three to five.

The novelty of this approach is its ability to prioritize a location's HSE risks using an algorithm containing natural language processing techniques. This cognitive tool treats reported HSE information as data to identify and flag priority HSE risks factoring in the frequency of similar reports and their associated severity

levels. The proof of concept has demonstrated the potential ability of the tool. The next stage would be to test predictive capabilities for injury prevention.

Introduction

As part of an ongoing focus of health, safety and environment (HSE) performance improvements, this paper describes how an oilfield services company aims to further learn from past HSE events to prevent recurrence of incidents using a cognitive HSE risk prediction and notification tool based on natural language processing. Robust HSE management with a proactive approach is essential across the industy to prevent incidents with a particular focus on preventing loss of life. Central to the management of HSE is the company's global HSE reporting system. This system gets hundreds of thousands of reports every month on various issues that affect the company's local population. In 2020, over 128,000 of the company's workforce used this system with over 11 million logins registered. The system processed more than 5.8 million HSE items including risk identification reports, meetings, audits, exemptions, management of change requests, observations, inspections, events, suggestions, and recognitions. There were, on average 5,300 proactive and reactive risk identification reports reported daily in 2020, which is over three reports every minute. This reporting system allows the company to search and review over two decades worth of reports and data. This provides significant opportunities for learning from events and trending, for benchmarking, and for the identification of emerging hazards. Automating the prioritization and dissemination of this magnitude of data is benefited from digital enablement to improve the efficiency of analysis.

A cognitive HSE risk prediction and notification tool based on natural language processing was developed to ease the data analysis efforts and significantly increase the value from the data within the global HSE reporting system. The ability to derive meaning from textual reports and automatically convert into actionable insights would be of significant benefit in obtaining operational HSE insights. The application of such an algorithm enables vast quantities of data in the form of text to be rapidly analyzed to provide trends and patterns which managers can use to make informed decisions to improve HSE performance.

To further raise awareness of this resulting information, a notification tool with a user interface has been developed for managers to view the prioritized risks to enable automatic notification for predetermined users with information relevant to their location.

Theory and Definitions

Natural language processing

The term natural language processing (Jackson, 2002) is normally used to describe the function of software or hardware components in a computer system which analyze or synthesize spoken or written language. The 'natural' aspect of the language processing is meant to distinguish human speech and writing from more formal languages, such as mathematical or logical notations, or computer languages, such as Java, LISP, and C++. Natural language understanding (NLU) is associated with the more ambitious goal of having a computer system actually comprehend natural language as a human being could.

A knowledge management system clusters an undifferentiated collection of text, for example in reports, documents or email messages, into a set of mutually exclusive categories.

Vectorization

It is a distributed representation of words in vector space (Mikolov, 2013).

Simple Mail Transfer or Transport Protocol (SMTP)

This is a transportation protocol (Sureswaran, 2009) used to transfer e-mail messages over the Internet. All e-mail servers use the SMTP to send e-mails from one e-mail server to another. SMTP is also used to send e-mail messages from e-mail clients to e-mail servers.

Clustering

Sorting an undifferentiated collection of objects into a set of mutually exclusive categories.

Clustering (Duda, 2012) procedures yield a data description in terms of clusters or groups of data points that possess strong internal similarities. Formal clustering procedures use a criterion function, such as the sum of the squared distances from the cluster centers and seek the grouping that extremizes the criterion function.

Description and Application of Equipment and Processes

Cognitive HSE Risk Prediction Algorithm

The pilot study used a selection of HSE reports (RIRs) submitted to the company's HSE reporting system over a 10-year time period as a reference dataset. A series of prediction algorithms were developed to analyze the dataset and categorize the HSE risks by potential severity and frequency using a formulated risk index methodology.

The process of developing the algorithm is shown in Fig. 1. Each of these processed steps shown were subject to continual improvement and analysis as described in more detail below.

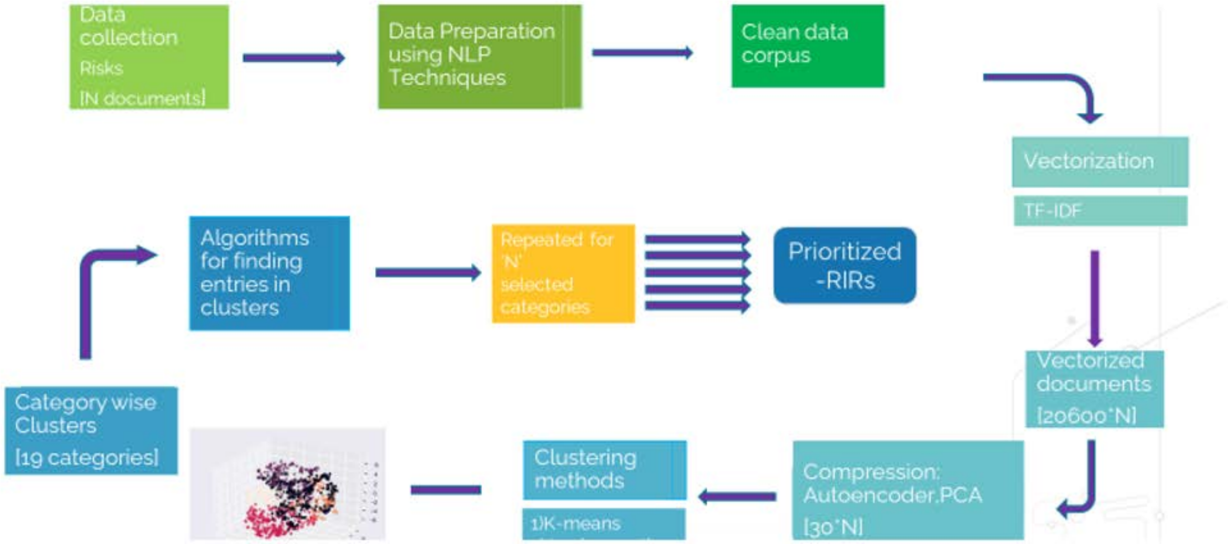


Figure 1—Cognitive HSE risk prediction algorithm flow diagram.

Data collection

For the pilot study, data was exported from the compay's global HSE reporting system. The selected dataset was from operations in the North America region. An initial set of over 7,000 HSE reports was used in the pilot sample dataset. An English language speaking geographic location was chosen to enable a sample dataset including diverse risks in several hazard categories to train the algorithm using sufficient data diversity. The reports were all submitted in English, which enabled the first round of algorithm validation to be performed without the added complication of multiple language translation requirements.

Data preparation

The data used in the pilot study contained several information fields including:

1. Unique report identification number
2. Incident date
3. Potential severity of loss consequence
4. Report summary

5. Report details.

Field 1 **Report identification number** is a unique report identifier used to reference the report; however, this data is not used in the analysis performed by the model.

Field 2 **Incident date** is an important parameter which is used to categorize the reports based on the date on which the event occurred, primarily used for the communication module within the tool.

Field 3 **Potential severity** is a numerical parameter used to capture the reporter's perception of the potential loss outcome of the event, based on the company's hazard analysis and risk control matrix methodology. This potential severity classification is guided by a standardized set of questions embedded within the global online reporting system to obtain consistency in classification.

Fields 4 **Report summary** and 5 **report details** contains a free text description of the event.

The summary has a brief description of the Report in less than 50 characters and report details field contains a description of the report in less than 4,000 characters.

Gaining insights from both the numerical parameters and non-numerical parameters such as the free text entry in both the report summary and report details field is the focus area of this approach.

The data collected from the HSE reporting system in a comma-separated values (csv) format had to be processed before analysis could be performed by the algorithm. This involved a sequence of processing stages using natural language processing as discussed below.

Tokenization

Tokenization is a process of converting the sensitive information to non-sensitive information by encryption. This is done to comply with the company and applicable regulatory data privacy requirements, the privacy of the users, and to protect any customer or company confidential information.

Tokenization works by converting plain text into a random token, to be used during processing by the algorithm.

Stop Words Removal

Since Fields 4 and 5 contain free text data entry, this contains lots of frequently used words, and for the purpose of data analysis of the text do not contribute any value.

An automated process was developed to remove a predetermined list of stop words from the free text data fields. Examples of the stop words removed include: 'the', 'a', 'to', 'where'.

Vectorization

The free text information contained in the HSE reports is in the form of sentences which needs to be made interpretable to machines, by assigning numerical/symbolic values to each word. The set of these values is known as a vector.

To vectorize the data, initially word-to-vec algorithms were considered. These are pre-trained models for vectorizing the data. Since the sentences used in the reports may contain technical HSE and/or operational terminology, it proved difficult to use a generalized algorithm for such scenarios. Re-training the word-to-vec models was not a practical solution as this is a time-consuming process which would have to have been repeated each time new terminology was introduced.

Based on the analysis of the experimentation conducted, the options for selection of a vectorization algorithm were narrowed down to term frequency–inverse document frequency (TF-TDF). The advantages of this method include not requiring using pre-trained models and that the algorithm worked effectively for the vocabulary used in the report submissions. The training of TF-TDF models is also less computationally expensive compared to other alternatives investigated.

In this approach, the 'tf' score was computed for all the words used in the analyzed report submissions, using the formula below.

$$tf(t,d) = 0.5 + 0.5\frac{f_{\{t,d\}}}{\max\left\{f_{\{t',d\}}:t'\in d\right\}}$$

$$idf(t,D) = \log_{N_i/\left|\left\{d\in D:t\in d\right\}\right|}$$

Here:
‘t’ is the term/word.
‘d’ is the document/report submission.
‘D’ refers to the set of documents/entire report submission dataset.

- The term frequency (tf) refers to the number of times the term ‘t’ occurs in the text of the document among all the occurrences of terms in the document ‘d’
- The inverse document frequency (IDF) refers to how commonly the term ‘t’ occurs within all the documents in the reference dataset (D)
- The ratio gives high values to the words frequently occurring in a certain document but are uncommon within the reference dataset (set of all documents).

Compression

Since a large quantity of data is generated after the vectorization, compression methods are needed to store and process the data efficiently. After vectorization, using TF-IDF, the data is made available in the form of a matrix of dimension ‘[Number of words * vector length]’. This matrix is quite large with the majority of the entries being 0, it is very sparsely populated with non-zero values.

Compression algorithms need to be applied for further data processing. For this stage, multiple compression algorithms were considered including the principal component analysis (PCA) which is a technique used for data compression ensuring information loss is minimal. Although this technique is proven to be useful for many applications, when used in this application a few drawbacks were identified which resulted in data loss due to improper compression.

An alternative approach using autoencoders was selected for data compression as these are widely used for dimensionality reduction applications.

Autoencoders are types of neural networks that learn efficient data encoding while decreasing the number of parameters required to store the data as shown in Fig. 2.

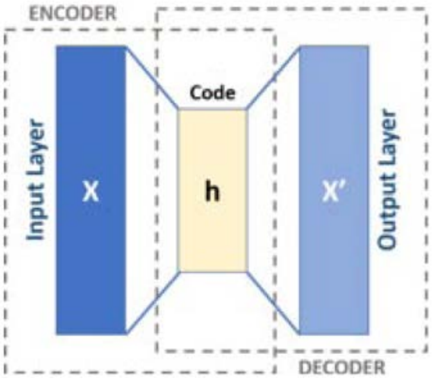


Figure 2—Encoder process.

The encoder processes the input data to generate the code which is significantly compressed to be more manageable and easier to process further. The code is then used to produce an output which is checked against the original input data to ensure minimal data loss in comparison to the input data.

The output layer is optimized to have the same data values as the input layer. However, having used the code, the data is available as a lower dimensional representation of the input and can be used as a proxy for the input.

The loss function is the difference between the input and the output and is calculated using the formula below.

$$L(x, x') = |x - x'|^2 = |x - \sigma'(W'(\sigma(Wx + b)) + b')|^2$$

Where x is the input, x’ is the output and W and b are the network parameters.

Clustering

Clustering is a machine learning technique which is used to group un-labelled data. Several types of algorithms were considered for the clustering process, such as hierarchical clustering, connectivity-based clustering and density models. Centroid-based clustering methods provided the best results for the pilot study data and provided a centroid for the cluster information, which is crucial for the methodology.

Based on the analysis and experimentation, the K-means clustering method was selected for this approach. In this methodology, the algorithms select n-centroids with the clusters being formed by the closeness of the data points to the centroids of the clusters.

The process is iteratively repeated by calculating the centroids in several stages, final centroids/clusters are reached when the distance of the points is minimized in relation to each centroid in the cluster. Compressed data can be shown in 3 dimensions or 2 dimensions, for the purposes of this pilot study, the data was plotted in 3 dimensions, as shown in Fig.3.

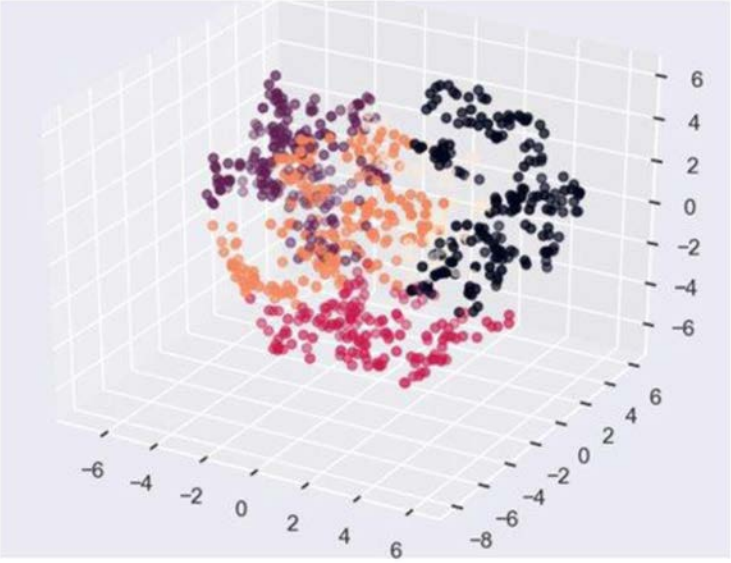


Figure 3—Category wise clusters based on severity and frequency of risk occurrence.

The centroids depict the data points which are most representative of the clusters in the dataset. The granularity of the data analysis is dependent on the number of clusters selected in the algorithm. A loss metric can be used to optimize the number of clusters identified within the dataset.

Fig.3 shows an example of the 3-dimensional representation of the data points with the distinct colors representing the different clusters.

The best entries from the cluster are identified as a representation for the cluster, this is done by finding the centroids of each cluster.

The centroid of a cluster has the minimum value for the following metrics, and this is captured as the tightness score or t. score.

$$\frac{1}{N}\Sigma_{i=1}\left|\left(x_i - x\sim\right)\right|$$

Here Xi is the vector corresponding to the ith entry of the cluster.
x~ is the centroid of the cluster.

The User Interface Module

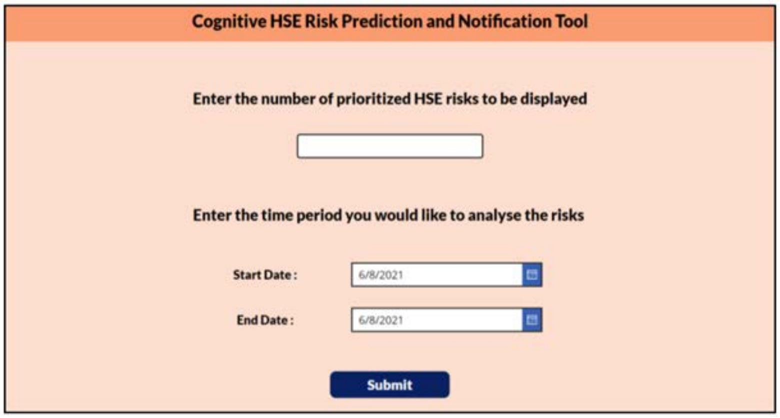
A user interface was developed to enable the user to specify a required number of HSE risks over a specific time period for a selected operational location.

To develop the user interface, two development platforms were considered:

1. A proprietary application development software
2. De-facto standard graphical user interface package for the programming language in which the cognitive HSE risk prediction algorithm was developed.

Option 1 proved to be not technically feasible. Therefore option 2 was selected for ease of module development. The user interface module captured the following details from the user, as shown in Fig. 4.

1. The number of prioritized HSE risks to be reported.
2. The time period to be used for analysis to generate the prioritized HSE risks.



Cognitive HSE Risk Prediction and Notification Tool

Enter the number of prioritized HSE risks to be displayed

Enter the time period you would like to analyse the risks

Start Date : 6/8/2021

End Date : 6/8/2021

Figure 4—Notification tool user interface template for users.

Notification Module

An email pipeline was set up to communicate the prioritized risks for a particular location to a group of users.

The module was designed to send notifications to the users using simple mail transfer protocol (SMTP). An SMTP object using smtplib was initially created to send the emails. A test group of receivers were also pre-defined for this pilot study.

By clicking on the submit button on the user interface, users were shown the list of prioritized HSE risks and the email pipeline was triggered to provide the information to the other users in the receiver group.

Results

The initial dataset was processed using the algorithm described above, over multiple iterations to identify the key improvements required for the various stages of this approach.

First Iteration:

- In the data cleaning stage, the algorithm successfully removed all the generic stop words (such as a, and, the)
- In the vectorization stage, using the Term Frequency - Inverse Document Frequency (TF-IDF) method the words in the HSE report texual descriptions were converted into vectors

- In the compression stage, an autoencoder removed the 'noise' from the input data
- In the agglomerative clustering stage, HSE reports with similar words were grouped into clusters with the number of clusters generated per category being dependent on the number of HSE reports. The output of this stage was used to generate a word cloud to visualize the frequency of word use in each cluster. Analysis of the generated word clouds indicated the initial list of stop words was not sufficient to clean the data and further processing was required to generate more meaningful word clusters.

Second Iteration:

- In the data cleaning stage, an additional list of more than 50 stop words such as 'place', 'go', 'put', 'used' and other common verbs and prepositions were removed from the word clouds
- In the subsequent agglomerative clustering stage, more meaningful clusters were obtained. Observations across the algorithm stages are described below.

Data cleaning stage

Preparing a collection of stop words is an iterative process, with multiple phases required to ensure that the most frequent words used in the HSE reports are not verbs that do not convey hazard identification information. The initial dataset analysis was used to train the algorithm to ignore stop words and words with less HSE significance, this required manual intervention. The effectiveness of this processing stage dictates the relevance of the output to the further stages of the algorithm analysis.

Compression stage

For each HSE report, the autoencoder was able to convert the the initial dataset representation of 20,500 neurons into a dense representation of only 30 neurons. This proved to be beneficial in computational efficiency.

Clustering stage

The sample dataset selected was extracted from the company's online HSE Reporting business system in assigned hazard categories. This enabled the algorithm to be used on the consolidated data and also on a more granular basis, by hazard category to determine any differences in the clustering results obtained. When the algorithm was run on a single HSE hazard category with the option to return five clusters from the analysis, the algorithm produced distinct clusters with the data points clearly clustered around each centroid. On a larger and more diverse dataset with various HSE hazard categories included, it was observed that selecting only five clusters to be returned was insufficient to clearly distinguish between each of the clusters. For more meaningful results the algorithm required the selection of a higher number of clusters, in the range of 20-30, to enable a clearer distinction between the clusters.

For HSE reports that had been reported with a higher severity category assigned, the algorithm was able to more effectively generate clusters to represent risks that had been more frequently reported within the dataset. For HSE reports that are more diverse in hazard category and severity as well as those which represented proactive reporting of HSE hazards, further development needs to be conducted to improve the effectiveness of the algorithm to prioritize the hazards and associated risks.

The algorithm took approximately five minutes to process 1,000 HSE risk reports, based on an average person's reading and comprehension speed, proved to be around 100 times faster.

It is recognized that the algorithm is limited by the language capabilities and access restrictions placed on some of the company's HSE reports, due to legal and confidentiality concerns. However, with the company's use of English as a global requirement for certain categories of HSE reporting and the quantity of HSE reports entered into the global online system, the benefits of using an algorithm are considered to outweigh these language restrictions.

Conclusions

Using standard software packages, a cognitive HSE risk prediction and notification tool has been developed, from algorithms to user interface. The results indicate that natural language processing techniques can be considered as a useful tool to analyze large volumes of HSE reports to produce insights based on the severity and frequency of hazards in the workplace. Further work is required to refine the data cleaning processes to enhance the benefits of such analysis tools and extend the initial pilot studies to analyze the most effective application of the algorithm to improve HSE perfomance.

References

R. Sureswaran, Hussein Al Bazar, O. Abouabdalla, Ahmad M. Manasrah, Homam El-Taj. 2009. Active e-mail system SMTP protocol monitoring algorithm. Presented at the 2nd IEEE International Conference on Broadband Network & Multimedia Technology.

Peter Jackson, Isabelle Moulinier, 2002. Natural language processing for online applications_text retrieval, extraction and categorization-John Benjamins Publishing Co.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems

Richard O. Duda, Peter E. Hart, David G. Stork. 2012. Pattern classification. Book published by John Wiley & Sons.