Table of Content

1. Introduction

This report documents the work that was done so far for the research titled “Natural Language Processing of Safety Reports in Nuclear Plants and Aviation,” supervised by Professor Najmedin Meshkati and Professor Yolanda Gil. All the work was performed during the Spring 2023 semester as part of CKIDS DataFest. The following students have contributed to the project:

Graduate Student Members: Anood Alkatheeri (MS ADS), Yassamin Neshatvar (MS EE), Samarth Saxena (MS ADS), Shruthakeerthy Srinivasan (MS CS), Rushit Girish Jain (MS CS), Abhishek Phalak (MS CS), Vaidehi Vatsaraj (MS CS).
Undergraduate Student Members: Iris Gordo, Shelby Wu.

The report first discusses the motivation (Section 2) for the project and the main objective (Section 3). Data sources are then presented in Section 4 (with access links) followed by the general approach (Section 5) to achieve the project objectives. Section 6 describes the baseline/test dataset (known as “Gold Standard”) while Section 7 presents the different algorithms that were explored to decide on which one to choose for the initial prototype. The details for data pre-processing are presented in Section 8, followed by a description of the proposed model (including training and testing) in Section 9. The initial results are shown in Section 10 where they are discussed in detail in Section 11. Finally, Sections 12 and 13 present the conclusion and future work, respectively.

2. Motivation

The Chernobyl Nuclear Power Plant disaster that occurred in Ukraine in 1986 was not only caused by a flawed reactor design that led to a catastrophic explosion. The plant's operators who were responsible for conducting a test on the reactor made several critical mistakes, including disabling key safety systems and ignoring warning signs. Similarly, at the Three Mile Island Nuclear Station, the operators on duty during the 1979 accident did not fully understand the plant's systems and were not adequately trained to handle the situation. At Japan's Fukushima Daiichi power plant, even though a massive tsunami led to the meltdown of several reactors in 2011, the plant's owner had a history of safety violations and a culture of complacency, where safety concerns were ignored or downplayed.

As Dr. Allison M. Macfarlane, who was the former chairman of Nuclear Regulatory Commission (NRC), once said, “There are many lessons that we must all take away from the accident at

Fukushima, but some of the most valuable extend beyond the technical aspects and are embedded in human and organizational behaviors. Among these is safety culture". In each of these cases, the safety culture of the organizations involved played a significant role in contributing to the nuclear disasters. Failure to prioritize safety, inadequate training and preparation, and a culture of complacency or secrecy all contributed to the accidents. The incidents might have been avoided or made less severe if these issues had been identified, confronted, and remedied.

In this project, we analyze safety reports from the Diablo Canyon Nuclear Power Plant (DCPP) in California, USA. Every year, the Diablo Canyon Independent Safety Committee (DCISC) publishes an annual report on the safety of Diablo Canyon. Analyzing these reports is crucial for understanding the safety culture of the plant and identifying safety problems, which in turn can help in taking the necessary actions to prevent future disasters. However, the issue with these reports is that they are too long (often over 900 pages), and there are currently no existing methods that can help organize this information. Thus, we propose to develop automatic methods that can help extract valuable insights about the plant's safety culture.

The methods developed as part of this research can apply to any safety-critical industries where failures are considered consequential, such as Nuclear and Aviation industries. Moreover, the results of this work could also be applicable to other power plants as it is expected that there would be more nuclear plants in the future due to climate change and the need for clean energy.

3. Objective

The project aims to utilize natural language processing (NLP) techniques to develop a model that can be used to extract "valuable" information from the Diablo Canyon Independent Safety Committee (DCISC) Annual Reports. The model will be trained using a dataset containing various incidents extracted from the report to relate them to INPO/NRC "Traits of a Healthy Nuclear Safety Culture." The goal is to analyze each extracted text from a given report to identify one or more associated safety traits. Below is an example of the desired output.

> **Paragraph**: Equipment problems due to aging have led to an increasingly negative trend in the station's Deficient Critical Component Backlog Orders. DCPP's performance on reducing or eliminating Safety System Functional Failures has not improved despite implementing a corrective action plan. - DCISC 24th Annual Report (2014)
>
> **Page:** 12
>
> **Associated Safety Traits:** Work Processes, Problem Identification, and Resolution

4. Data Sources

Two main data sources were used in this project, which are described as follows.

4.1. Diablo Canyon Independent Safety Committee (DCISC) Annual Reports.

Link to reports: https://www.dcisc.org/annual-reports/

The Diablo Canyon Independent Safety Committee (DCISC) is a three-person Committee charged with reviewing and making recommendations concerning the safety of operations at Pacific Gas and Electric Company's (PG&E) Diablo Canyon Nuclear Power Plant (DCPP). There are currently 12 annual reports available electronically on the DCISC website (from 2010 to 2022).

Each annual report is divided into two volumes: Volume I and Volume II. Volume I includes a summary of committee activities and documents received by the DCISC during the reporting period. Volume II contains a list of documents received by the DCISC, public meeting notices and agendas, an operations summary for the reporting, information brochure, and minutes.

What makes this data challenging is it exists as raw unstructured text in a PDF format, which also includes tables and images. Moreover, the text sometimes has spelling errors and is missing punctuations. The length of each PDF file is often around 900 pages, but there is a lot of repeated information throughout the report. Thus, appropriate Python libraries are needed to deal with such long PDF files and to extract the necessary text only in a clean way.

4.2. “Traits of a Healthy Nuclear Safety Culture” document developed by the Institute of Nuclear Power Operations (INPO).

Link to document:
https://www.energy-northwest.com/Documents/Employee%20Portal/INPO%20Safety%20Culture%20Traits.pdf

This document describes the essential traits and attributes of a healthy nuclear safety culture to create open discussion and evaluate the continuous evaluation of safety culture in the nuclear energy industry. It has been translated into many languages and is used by many nuclear power plants worldwide. The 10 safety traits described in the document are as follows:

— Personal Accountability
— Questioning Attitude
— Effective Safety Communication

— Leadership Safety Values and Actions

— Decision-Making

— Respectful Work Environment

— Continuous Learning

— Problem Identification and Resolution

— Environment for Raising Concerns

— Work Processes

What is challenging about this document is it also exists as unstructured text data in a PDF format that must be pre-processed appropriately in Python before use. Unlike the first source, the document is not very long (around 38 pages). However, there is some overlap among the 10 safety traits in terms of their attributes description and behavior examples. This would make it challenging for the model to distinguish between them.

5. General Approach

Below is a general set of objectives that the team has defined in order to approach this project.

- The model should identify valuable paragraphs from DCISC reports that describe a safety incident and root cause evaluation (if available).
- The model should analyze paragraphs to identify/predict the associated safety traits using the list of 10 safety traits from the “Traits of a Healthy Nuclear Safety Culture” booklet of INPO.
- The team should prepare a “gold standard” document by manually reading, analyzing, and extracting valuable paragraphs from different DCISC reports and labeling them with the most relevant safety traits. This “gold standard” will be used to evaluate the model's performance.
- The model should predict the most relevant 4 safety traits per incident/paragraph based on some defined threshold/probability.
- The model’s accuracy should be evaluated using a metric that is suitable for multi-label predictions (ex: F1-score, Hamming score/distance, Jaccard score/distance…)
- The team should create a visualization depicting the attributions of each extracted incident/paragraph with their relevant safety traits. This visualization should emphasize the most commonly attributed safety traits.
- The model should be used on a series of DCISC reports to understand the safety culture trend over time in Diablo Canyon Power Plant.

6. “Gold Standard” Dataset:

In order to have a baseline standard against which we evaluate the model’s performance, the team prepared a “Gold Standard” dataset [1]. This dataset consists of manually extracted paragraphs from several DCISC reports that describe a safety issue along with root cause evaluation (if available). Each paragraph was manually labeled with a list of associated safety traits based on reading the INPO booklet and using keywords. Moreover, the labels were further verified by having multiple reviewers. So far, around 67 paragraphs have been labeled, which will serve as the test set for our model.

7. Exploration of Algorithms

There are so many existing solutions that tackle similar problems of identifying valuable texts and classifying them into one or more target labels. As mentioned in the “Project Objectives” section, there are two main tasks to be achieved:

- Identifying valuable paragraphs from DCISC reports that describe a safety incident and root cause evaluation (if available).
- Analyzing the extracted paragraphs to identify/predict the associated safety traits using the list of 10 safety traits from the “Traits of a Healthy Nuclear Safety Culture” booklet of INPO.

This section describes the approaches that were explored to achieve each task.

7.1. Pre-trained Sentiment Analysis Models

For the first task, the team determined that a sentiment analysis algorithm can be used, since the valuable paragraphs of interest that describe a safety incident will most likely have a negative sentiment. Due to the tight schedule of DataFest, several pre-trained sentiment analysis models were evaluated instead of training one from scratch. The pre-trained models were obtained from the website “Hugging Face”, which is an open-source community that allows users to deploy machine learning models. In order to evaluate the models, a test set was prepared which consists of sentences from the gold standard document (which are negative-toned) and 75 general positive-toned sentences taken from a DCISC report. The F1-score metric was used to evaluate the accuracy of the sentiment analysis models on the test set. A good sentiment analysis model should be able to accurately distinguish the negative-toned paragraphs from the positive-toned ones. Table 1 below summarizes the comparison results.

Table 1: comparing performance of sentiment analysis models

Model	FinBERT	BERTweet	Siebert/roberta-large-english	Sbcbi/sentiment_analysis	Amansolanki/autoNLP-tweet
F1-Score	91%	87%	87%	86%	86%
Model	Transformer Library's Default Model	Seethal/generic_dataset	Avichr/heBERT	Liyuan/amazon_review	Cointegrated/rubert-tiny
F1-Score	83%	79%	72%	68%	65%

As can be seen in the table, the “FinBERT” sentiment analysis model was found to be the most accurate. Thus, it can be used in the model to evaluate the “negativeness” of paragraphs in order to only focus on analyzing the relevant ones that discuss safety culture problems.

To achieve the second task, the team wanted to split up and explore/evaluate different approaches to see how they can handle the problem. A summary of these approaches along with their pros and cons are discussed in the following subsections.

7.2. Unsupervised Rule-Based Approach

The first approach explored was a rule-based approach to perform aspect-based sentiment analysis (ABSA). ABSA works by first identifying the aspects that are discussed in a text and then performing sentiment classification for each aspect to classify it as positive or negative. Aspect identification can be done by analyzing the sentence structure and the grammatical dependencies between words, which was done using Python’s spaCy library. Figure 1 shows the type of visualization that can be generated by spaCy. In this example, the main aspects that are described by adjectives are “communication” and “leadership”. For each of the 10 safety traits, a list of associated words can be generated by reading the traits description in the INPO document. The aspect “communication” could be related to the safety trait “Effective Safety Communication” while the aspect “leadership” is associated with “Leadership Safety Values and Actions”. Both of these aspects are described by negative adjectives, “communication” is described as “poor” while “leadership” is described by “lack”.

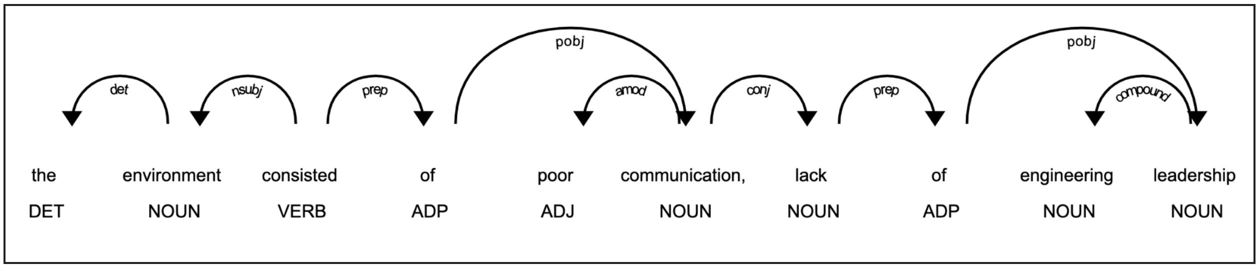

Figure 1: Dependency visualizer generated by spaCy

Thus, given a paragraph, this approach can extract aspect-adjective pairs and a sentiment analysis model can be used to filter those pairs to only the ones with negative adjectives. The aspects can then be mapped to their associated safety traits. The advantages of this approach is that it is completely unsupervised and does not require the use of training data. However, there are several problems with this approach. First, a lot of rules have to be manually defined in order to extract the aspect-adjective pairs for different types of sentence structures. Sometimes a paragraph could be implicitly discussing a safety culture problem and it may not be clearly mentioned as an aspect-adjective pair. Another challenge is that the aspects can sometimes consist of multiple words and not just a single one. Thus, the algorithm should be able to handle the extraction of multi-word aspects.

7.3. Semi-Supervised Approach

Semi-supervised approach, as the name suggests, is a training procedure that combines the core ideas of both supervised and unsupervised approaches. Basically, it allows training of the model on labeled data with lesser manual effort in labeling than what is required for a supervised training. We begin with manual labeling of a balanced subset of our unlabelled dataset. This subset is referred to as the “seed data” for the semi-supervised training. Thereafter, a model is trained using the seed data. The resulting model is used to label another subset (again balanced) of the unlabelled dataset. The cycle continues. On a very superficial level, it could be understood as an approach that involves a cycle of training and labeling. We iterate upon these two steps until we get a satisfactorily accurate model or if the dataset is completely labeled.

The base model used here was a BERT model. A subset of 1200 sentences was picked randomly from the dataset. After manual labeling, initial training was performed with 1000 sentences. The other 200 were used for testing. The test-train ratio was maintained at 1:5 for all subsequent iterations. The result was a 45.3% accurate model. This was below the expectations. Tuning the model's hyperparameters led to a marginally increased accuracy of 48%. This model was used to train another 1200 sentences. In total, 7 iterations were performed with the final accuracy at 43%, which was below the method we chose later.

There are several improvements that can be done to get this method to train a much better model. The two main challenges encountered during the implementation include: -

- *Catastrophic Forgetting*: Testing in later iterations (6th and 7th) suggest the model was forgetting what it had learned in the previous iterations.
- *Balance in Dataset*: Although we tried maintaining balance in the dataset, it was evident that some classes were not being identified with as much accuracy as the others were being identified with. Instances of some classes were found to be much lesser than some other more common classes in the test set. However, the gold standard dataset was not prepared until then.

7.4. Unsupervised Classification Approach

Another unsupervised approach that does not require labeled training data is classification based on the Lbl2Vec algorithm [2]. The key idea of Lbl2Vec is to first represent each category (or safety trait) by a set of keywords. Then, a paragraph of interest can be represented as an embedding vector based on the words it contains. Similarly, the classification categories (labels) can also be represented as vectors. The algorithm then classifies each paragraph to its category by computing the similarity between the vectors (cosine similarity). However, the issue with this approach is that each paragraph will only be classified to one category since it does not allow overlap. In the case of this project, a safety incident can be attributed to one or more associated safety traits.

7.5. Unsupervised Clustering Approach

To allow the attribution of a text to multiple categories, a probabilistic text clustering algorithm can be used (soft clustering). One technique is topic modeling using Latent Dirichlet Allocation (LDA), which is an unsupervised statistical approach that identifies the topics discussed in a given text. The advantage of this approach is that it would be able to assign multiple safety traits to a given incident/paragraph (it allows overlapping). Generally, the way that LDA automatically discovers topics is by finding the mixture of words that is associated with each topic. However, in the case of this project, the "topics" do not need to be discovered since they are already given as the 10 safety traits. Thus, a modification to the LDA algorithm must be made to allow the model to converge to the existing topics.

Luckily, a developer online created the “GuidedLDA” model, also known as “SeededLDA”, which is a semi-supervised learning algorithm [3]. The idea is to set some seed words for topics that the user believes are representative of the underlying topics in the corpus and guide the model to converge around those terms. In this case, the topics correspond to the 10 safety traits as defined by INPO. Examples of seed words that can be used per safety trait are as follows:

Personal Accountability: responsibility, accountability, help, support, trained, qualified, understand, complete, involvement

Questioning Attitude: complacency, complacent, challenge, error, hazard, caution, discrepancy, anomaly, assumption, question, uncertain, unknown, risk, trend, unexpected, unclear, degrading, aging

Effective Safety Communication: communication, licensee, event, report, documentation, request, LER, information, safety, prompt, share, respond, listen, concern, expectation, clear

Leadership Safety Values and Actions: leadership, management, leader, owner, ownership, program, guidance, policy, resource, staffing, oversight, reinforce, priority, plan, delegate, align, define, manage, resolve, address, translate, funding, implementation, violation

Decision Making: thorough, conservative, systematic, consistent, process, choice, consequence, authority, future, timely, executive, senior

Respectful Work Environment: trust, respect, opinion, dignity, fair, disagree, receptive, valuable, tolerate, value, insight, perspective, collaboration, conflict, listening

Continuous Learning: learn, training, assessment, improve, performance, scrutiny, monitor, adopt, idea, benchmarking, knowledge, competent, skills, develop, acquire

Problem Identification and Resolution: identify, corrective, action, issue, yellow, red, prevent, foreign, poor, inadequate, degraded, evaluation, problem, cause, root, investigation, investigate, recommendation, resolution, mitigate

Environment for Raising Concerns: environment, fear, harassment, discrimination, promote, severity, failure, submit, report, expired, raise

Work Processes: engineering, control, activity, contingency, production, schedule, work, margin, operate, maintain, maintenance, procedure, package, accurate, current, backlog, instruction, operation, design, requirement, standard

After the model is trained, it can be used to predict what safety traits are associated with a given paragraph describing an incident. The advantages of this model is that it can classify a paragraph to multiple safety traits, not just one. This is because the output of the model will be a list of probabilities that a paragraph is discussing each of the safety traits. Thus, a threshold can be defined to assign a safety trait to the paragraph based on its probability value. However, a limitation of this model is it might be highly dependent on the list of seed words defined, and no words can overlap between the different categories. Another challenge is that the probability threshold must be reasonably defined in order to minimize incorrect predictions.

After evaluating the pros and cons of the different approaches above, the team decided to continue with the “GuidedLDA” model for the first prototype. The details of the model training and prediction algorithm is described in Section 8 “Proposed Model”.

8. Data Pre-Processing

Before processing the PDF files of DCISC reports, it was important to manually scan through the report first in order to determine which general headings/sections contain relevant information. Although the report is very long (around 900 pages), many of the pages contain information about the safety committee, their evaluation process, and how meetings are conducted. This kind of information is generally irrelevant to the task at hand since we are more interested in the safety culture of the power plant and any incidents that have occurred. Thus, the following report sections were determined to contain the relevant information:

Volume I - Main Report

- 1.0 Introduction
    - 1.4. Committee Member Site Inspection Tours and Fact-finding Meetings
    - 1.5. Visits by DCISC Members to California State Agencies, Outreach Activities….
- 4.0 Summary of Major DCISC Review Topics

Volume II - Exhibits

- B.3 Minutes of [DATE 1] Public Meeting
- B.6 Minutes of [DATE 2] Public Meetings
- B.9 Minutes of [DATE 3] Public Meetings
- C. Diablo Canyon Operations
- F. DCISC Open Items List
- I. Current and Past DCISC Recommendations and PG&E Responses

The PDF files of the report were then processed using Python’s “pdfplumber” library in order to extract the text elements page-by-page, focusing on the headings of interest. The page numbers corresponding to the headings of interest were also extracted. For each heading, the texts from multiple pages were combined into a single text and then split into paragraphs. As part of the text cleaning process, numbers, punctuations, acronyms, and stop words were removed. Lowerization (transforming letters to lower-case) as well as lemmatization (replacing words with their root form) were also performed. Furthermore, month names and people names were also removed. The final output of preprocessing is a list of cleaned paragraphs (from the headings of interest) that is ready to be used as training data for the model.

9. Proposed Model

As explained in the section “Exploration of Algorithms”, the team decided to go for the unsupervised clustering approach to train the initial prototype. Specifically, the approach chosen is the topic modeling algorithm “GuidedLDA” or “SeededLDA”. The model training and prediction algorithm is described as follows:

9.1. Model Training

The training data is a list of cleaned paragraphs that are extracted from the headings of interest. For the initial prototype, only one DCISC report was used (31st annual report, 2021). For the GuidedLDA model, the training data must be converted to a document-term sparse matrix. The rows of the matrix correspond to the paragraphs while the columns represent all the unique vocabulary. For each paragraph, the row values can either consist of a simple frequency count of the words that exist in the paragraph or a weighted count using TF-IDF.

An important aspect of the training is defining the “seed words” for each of the 10 safety traits, which will guide the model towards convergence. The same seed words defined in section 6.4 “Unsupervised Clustering Approach” were used in the training, which were lemmatized to their root form and then mapped to their category id (each safety trait is represented by a digit from 0 to 9).

The GuidedLDA model is then fitted using the document-term matrix as well as the seed words to guide the training. In order to control how biased the LDA model should be towards the seed words, a value for the “seed confidence” can be specified, which ranges from 0 to 1. A seed confidence of 0.9 means that the model will be biased by 90% more towards the seeded topics.

As a result of training the model, a word-topic distribution can be generated, which consists of a list of words that are frequently associated with each topic or safety trait based on the training paragraphs. A sample output of the top words per topic as generated by the GuidedLDA model can be seen in a word cloud illustration as shown in Figure 2. This illustration helps users understand what are the most important words associated with each safety trait, where the size of the word represents its frequency of occurrence. It also helps developers assess whether the algorithm has successfully captured the topics.

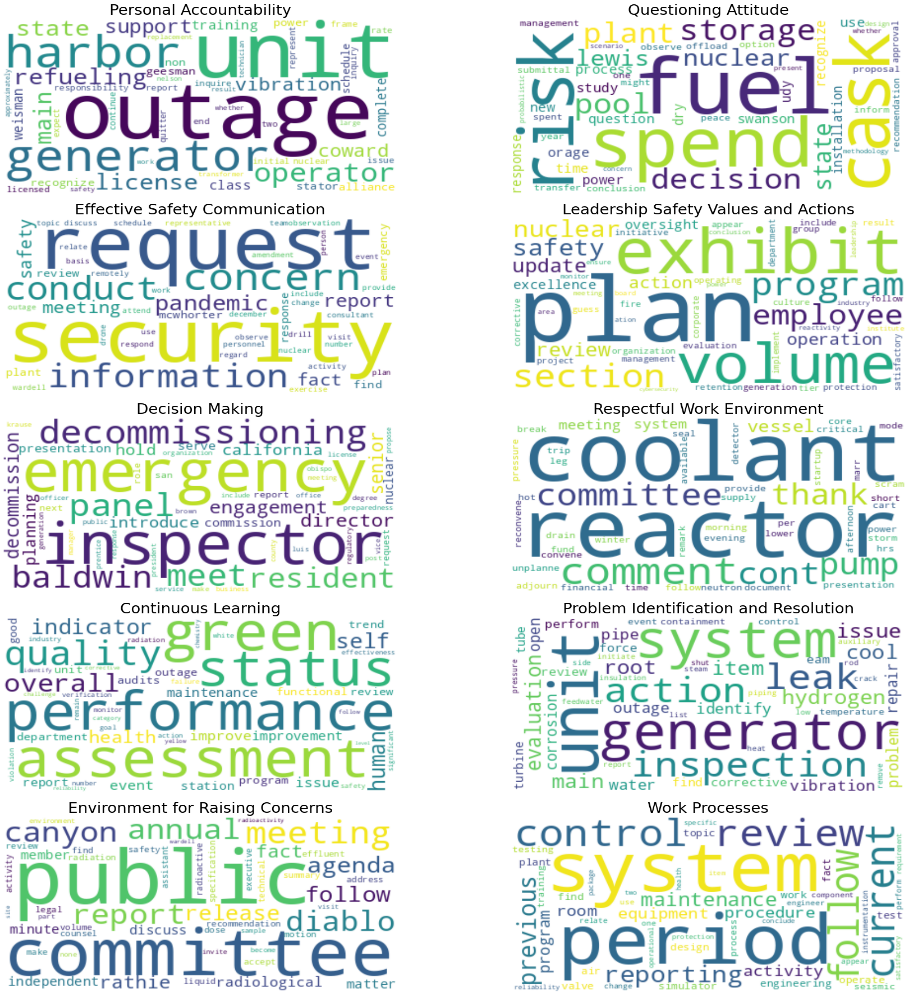

Figure 2: World cloud for top words per topic

9.2. Model Testing

Given a paragraph, the model will output a list of probabilities that each word belongs to each of the 10 safety traits. These probabilities can be aggregated (averaged) across all words in the paragraph. The aggregated probabilities will then be used to predict the safety traits that are discussed in the paragraph. For instance, if the aggregated probability for a certain safety trait is higher than 10%, it is likely that the paragraph is discussing this trait. In order to test how well the model can predict the safety traits discussed, the “Gold standard” dataset can be used, which contains manually labeled paragraphs. An example of how the GuidedLDA model can be used to predict the associated safety traits is as follows:

Paragraph: “Equipment problems due to aging have led to an increasingly negative trend in the station’s Deficient Critical Component Backlog Orders. DCPP’s performance on reducing or eliminating Safety System Functional Failures has not improved despite implementation of a corrective action plan.” - DCISC 24th Annual Report (2014)

Aggregated Probabilities:
Personal Accountability: 0.0476
Questioning Attitude: 0.0402
Effective Safety Communication: 0.0448
Leadership Safety Values and Actions: 0.1087
Decision Making: 0.0056
Respectful Work Environment: 0.101
Continuous Learning: 0.3046
Problem Identification and Resolution: 0.1412
Environment for Raising Concerns: 0.0248
Work Processes: 0.1815

Predicted Traits: ['continuous learning', 'leadership safety values and actions', 'problem identification and resolution', 'work processes']

Actual Traits: ['problem identification and resolution', 'work processes']

As can be seen in the example, if we use the 10% probability threshold assumption, the model outputs 4 likely traits. However, only two of them are consistent with the actual traits that were manually determined in the “Gold standard” dataset.

10. Initial Results

The trained GuidedLDA model was then used on the entire test dataset to predict the associated safety traits for each paragraph. To quantify the accuracy of the model and its performance, a necessary exploration of the evaluation metrics is needed in order to determine which suitable metrics to use.

10.1. Evaluation Metrics

Since the nature of the classification problem is multi-class and multi-label, there are 4 possible metrics that can be used: Accuracy, precision, recall, and F1-score. A brief description for each of the metrics is given as follows.

- **Accuracy/Hamming score:** Proportion of correctly predicted labels to the total number of labels (predicted and actual) for that instance.
- **Precision**: Proportion of correctly predicted labels out of all predicted labels.
- **Recall**: Proportion of correctly identified labels out of all actual labels.
- **F1-Score**: Harmonic mean of precision and recall, gives equal weightage between precision and recall.

The four metrics will be evaluated for each test paragraph, then the overall model performance can be evaluated by calculating the average values of the scores across all paragraphs. An example of how the scores are evaluated for each test paragraph is shown as follows.

Paragraph: RC2: Maintenance leadership has not been proactive in its approach to shortfalls in human performance standards and use, including the failures to consistently perform task previews and establish clear standards for work order use and adherence.
Predicted Traits: ['continuous learning', 'effective safety communication', 'leadership safety values and actions', 'work processes']
Actual Traits: ['continuous learning', 'work processes']

Accuracy = $Correct\,predictions/(Predicted\,labels\,+\,Actual\,labels)\,=\,2/4\,=50\%$
Precision = $Correct\,predictions/Predicted\,labels\,=\,2/4\,=\,50\%$
Recall = $Correct\,predictions/Actual\,labels\,=\,2/2\,=\,100\%$
F1-Score = $2\,*\,Precision\,*\,Recall\,/(Precision\,+\,Recall)\,=\,66.67\%$

10.2. Model Performance

In order to evaluate the overall performance of the initial prototype, the “Gold Standard” Dataset can be used as the test set. For each test paragraph, the model predicts the safety traits that are discussed. Then, the accuracy of the prediction can be evaluated using the metrics mentioned in the previous section. Finally, the metrics are averaged for all the paragraphs to obtain overall model scores. However, the assumption used here is that the safety traits that are manually determined in the “Gold Standard” document are assumed to be “correct”. In reality, since the traits were manually determined by reviewers, there might be some subjectivity involved. Nevertheless, the “Gold Standard” can still be used for the purposes of first prototype evaluation.

Baseline Model

In addition to the “Gold Standard”, a baseline model can also be used to evaluate the performance of the prototype model. The baseline model used is a trivial classifier that randomly picks an output label, with probability given by the label’s frequency of occurrence in the training data. This baseline model will help determine if the current prototype performs better than random or not and, if so, by how much. The F1-score of the baseline model (using “Gold Standard” as the test set) is around **30%**.

Results

The following Table 2 summarizes the results of the model in terms of the evaluation metrics and how it compares with the baseline model. Generally, as it can be seen, the initial prototype based on the “Guided LDA” algorithm performs better than the random baseline model, although not by a sufficiently large margin. The current model is able to successfully detect some of the safety traits discussed in a given incident paragraph. However, there is still a lot of room for improvement as discussed in the “Future Work” section (Section 13).

Table 2: Model results on “Gold Standard” dataset

	Accuracy	Precision	Recall	F1-Score
Baseline Model	20%	39%	25%	30%
Prototype Model	36%	46%	54%	48%

11. Discussion

Although the initial model performance is not as exceptional as the team had hoped, it can still be used to generate useful insights about the safety culture of the Diablo Canyon Power Plant. For instance, a table can be generated which summarizes the most negative texts from the DCISC report, which could represent the safety incidents that have occurred during the reporting period. This can be done using the selected pre-trained sentiment analysis model (Section 7.1). Additionally, for each incident, the list of associated safety traits can also be shown, which were predicted by the initial model. A sample output of this table can be seen in Table 3. The paragraphs are from the DCISC 31st annual report that were used in the training process.

Table 3: “Negative” paragraphs with their predicted safety traits

Training Paragraph	Predicted Traits	Sentiment	Sentiment Score
DCPP's performance on the NRC's performance metrics placed DCPP in the highest performance category, Column 1 of the NRC's Licensee Response metric. There were four violations identified during 2020 compared to eight in 2019 and this represents the lowest annual total since 2016.	Continuous Learning	negative	0.973767
Operations developed a Plant Status Control Action Plan to address this performance decline which included a common cause evaluation, increased observations and communications, and a site-wide video to demonstrate strong component positioning behaviors. The failure to effectively address these challenges, including two Station Level Events SLEs that occurred the remainder of 2019, contributed to a yellow window for operations.	Continuous Learning, Leadership Safety Values and Actions, Problem Identification and Resolution, Effective Safety Communication	negative	0.967037
In 2017 the overall System Health was rated Yellow, due to component aging and parts obsolescence, and a compressor replacement plan had been initiated.	Work Processes, Continuous Learning, Problem Identification and Resolution	negative	0.963338
DCPP's Safety Fair was an excellent activity that encouraged employee awareness and knowledge of various important work safety topics in preparation for the upcoming outage. DCPP identified significant negative trends in Operations Department human performance during 2019.	Continuous Learning, Work Processes, Leadership Safety Values and Actions	negative	0.962032
Nuclear Instrumentation Detector Arrangement DCPP reported that the NIS is in good health. Two Unit 2 Intermediate Range Detectors required replacement in November 2019 and April 2020 due to abnormally high indications caused by faulty electrical connections. The NIS has been operating normally since then.	Continuous Learning, Problem Identification and Resolution, Personal Accountability	negative	0.952874
Other system problems occurring in the past included issues with backward rotation of idle fans, which could then trip upon starting due to high currents.	Problem Identification and Resolution, Work Processes	negative	0.947836
Dr. Lam commented that reports in the local media have alleged the problems with the Unit 2 Main Generator were the results of willful gross negligence by DCPP management.	Problem Identification and Resolution	negative	0.948661

Another valuable output of the model is a visualization that shows what were the most discussed safety traits in a given report. This would be a very helpful insight to the end users as it would highlight the most pressing safety culture issues in the power plant during the reporting period. For the 31st DCISC annual report (2021), a breakdown of the most discussed safety traits are shown in Figure 3, which were based only on the paragraphs with negative sentiment scores. As can be seen, most of the incidents were related to the “Problem Identification and Resolution” trait, followed by “Personal Accountability” and “Continuous Learning”. This highlights the need for the company employees to be more proactive in identifying potential problems early on and to take part in correcting them.

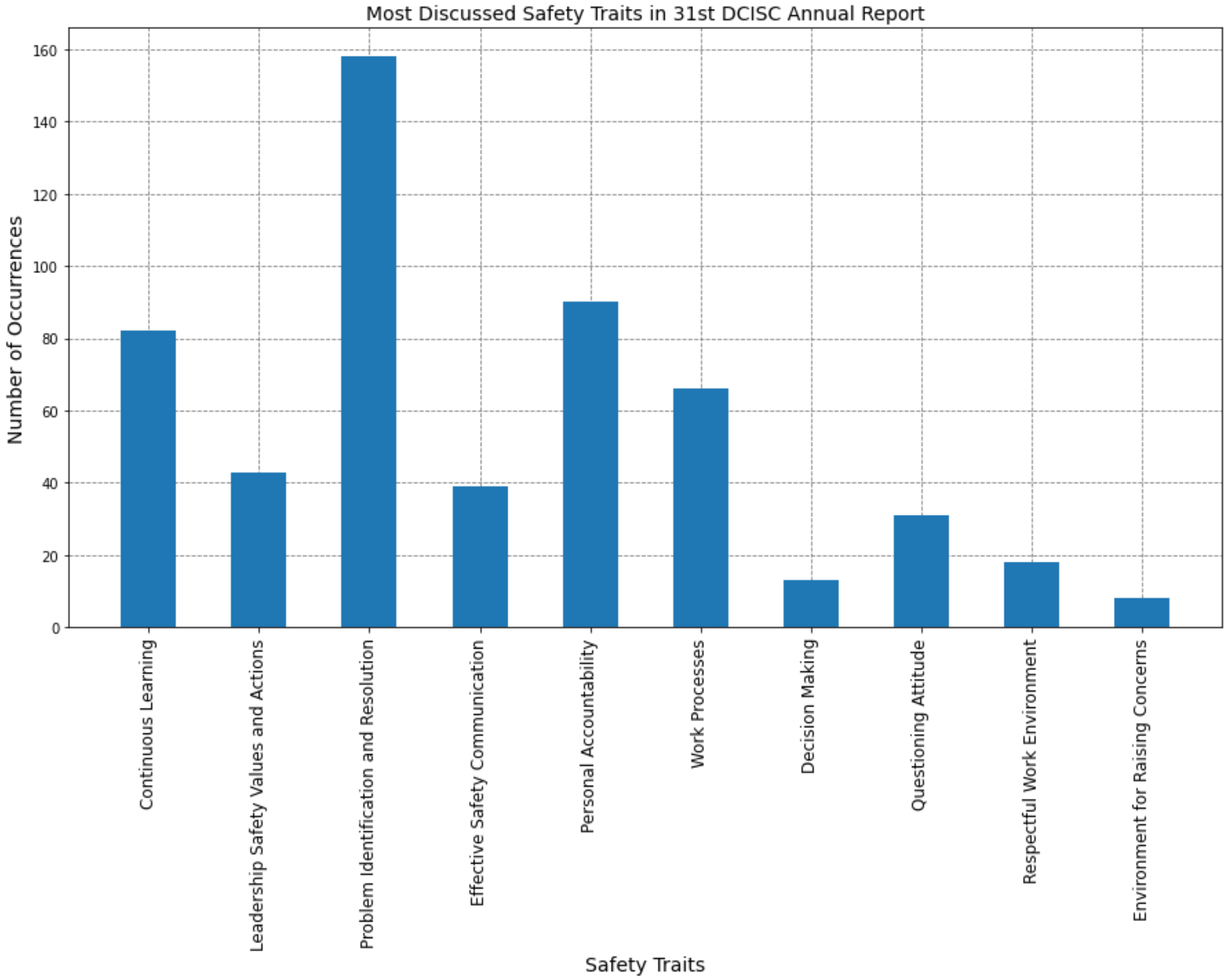

Figure 3: Bar chart showing the most discussed safety traits in DCISC 31st report

12. Conclusion

This project presents the development of an initial prototype model to analyze the safety reports produced by DCISC using natural language processing techniques. The proposed model extracts paragraphs from 31st DCISC annual report, determines the sentiment of the paragraph, and predicts the safety traits that are discussed. The assumption is that paragraphs with negative sentiments contain information about safety incidents. Thus, the model can be used to predict which of the 10 safety traits from the INPO booklet are responsible for the incidents in the power plant. The prediction accuracy of the model was evaluated by comparing against a “gold standard” as well as a baseline model. Moreover, visualizations were generated to display words associated with each safety trait (word cloud), summarize the incidents along with their associated safety traits in a table, and highlight the most discussed traits in a given report as a bar chart. Although the initial prototype performs better than random, there is still a lot of room for improvements as discussed in the next section.

13. Future Work

There are several improvements that can be done in the near future to allow the model to be as useful as possible. Ultimately, a dashboard can be created to summarize the safety-related incidents from a report along their associated safety traits in addition to general statistics about the most significant safety issues that need to be addressed. This would be a very helpful tool for managers and safety officers as it would give them an overview about the current safety culture and how it has changed over the years. The dashboard would enable effective communication of model findings to decision-makers to help resolve safety culture issues in the power plant. However, currently, the model is not equipped to handle multiple DCISC reports since they don’t all share the same outline structure. Thus, some changes in the pre-processing methods might be needed to allow the model to handle multiple DCISC annual reports for the trend study.

Moreover, it was determined that the performance of the model is highly dependent on the list of seed words used in the training of the “Guided LDA” algorithm. Thus, one way to improve the model predictions is to refine the list of seed words used for each safety trait by using an algorithm that will automatically extract the words from the INPO booklet. Another obvious improvement is to increase the size of the training data by utilizing multiple DCISC reports instead of just one. Increasing training data would improve the model’s ability to generalize and make accurate predictions on unseen or new examples.

In terms of how “Guided LDA” works, the current model can only determine the probability that a single word belongs to a safety trait category. To find out the probability that an entire paragraph discusses a certain trait, the current model manually calculates the average probabilities across all words. This may not be an ideal way to obtain an overall probability since it assumes that words are independent of each other. Therefore, it is necessary to explore different probability aggregation methods or even an entirely different algorithm in order to find ways to improve the prediction accuracy.

14. References

[1] "labelled_dataset." Google Docs, April 2023, https://docs.google.com/document/d/1BQuQv7p4j73t-XcdU0YIp0F2VNVHxGdvbR1SOsjbL7Y/edit

[2] Sethi, S., "Unsupervised Text Classification with LBL2Vec," in Towards Data Science, May 6, 2021. [Online]. Available: https://towardsdatascience.com/unsupervised-text-classification-with-lbl2vec-6c5e040354de

[3] "GuidedLDA Documentation," in Read the Docs, 2021. [Online]. Available: https://guidedlda.readthedocs.io/en/latest/