1. Summary of problem statement, data and findings.

Every good abstract describes briefly what was intended at the outset, and summarizes findings and implications.

It is not always easy to find databases from real-world manufacturing plants, and this data set is basically records of accidents from 12 different plants in 03 different countries which every line in the data is an occurrence of an accident.

Our dataset typically consists of the Location, Accident Level, Gender, Industry Sector and the Description of the Accident.

The companies want to understand why employees still suffer some injuries/accidents in their plants which even lead to their death. So they need our help to explore and take newer insights from the data.

Our objective is to build a Chatbot which can obtain the Description of the accident from the user as an input and then build a model to predict the Accident level based on which industries can take appropriate action.

In order to build a Chatbot we will only need the Description feature as our Independent Variable and the Accident Level as our Target Variable.

Findings :

- We have no missing values in this dataset.
- We have seen duplicate values in this dataset.
    - We have created a new feature by combining ‘Accident Level’ and Potential Accident Level’.
    - Target variable – ‘Custom Accident Level’ distribution is not uniform.
    - Data is highly imbalanced and so we have augmented and balanced the data for future use.
    - Contextual Word Embedding is used handling imbalanced dataset.

We will create another dataset without processing such as stemming and lemmatization, which could be used even in NLP models such as BERT. We will use the following Augmentation Techniques:

- Contextual Word Embedding Augmentation (ContextualWordEmbsAug)
- Synonym Augmentation (SynonymAug)
- Random word augmentation (RandomWordAug)

Since the aim is to build NLP based chatbot as described in the report, we drop all columns except the **Description** and the **Target Variable (Accident Level).**

2.Overview of the final process.

Briefly describe your problem methodology. Include information about the salient features of your data, data pre-processing steps, the algorithms you used and how you combined techniques.

Since the aim is to build NLP based chatbot as described in the report, we drop all columns except the **Description** and the **Target Variable (Accident Level).** Since the data was highly imbalanced, we have used different augmentation techniques to do so. We have created 2 datasets, one for BERT and one for the other models. We have also label encoded the data for better understanding.

Balanced dataset will provide better results from our model and help us in better analysis of the reason for accidents.

Salient Features of our data :

- We have no missing values in this dataset.
- We have seen duplicate values in this dataset.
  - Created a new feature by combining 'Accident Level' and Potential Accident Level'.
  - Balanced Data and Augmented data that will provide us better results from our models.

Pre-processing steps **:**

As a part of our pre-processing procedure, we have converted all characters to lower case, removed white space, removed special characters. Apart from these we must balance our Target Variable which is heavily imbalanced. The visualizations of which have been shown in the notebook.

In order to handle the imbalance, we have performed data augmentation and we have used various augmentation techniques like Contextual Word Embedding Augmentation, Random Word Insertion and Synonym Augmentation. Inside these augmentation techniques also we have used various sub-techniques like Insertion, Deletion, Substitution and Replacement in order to avoid redundancy, and to produce a balanced and augmented dataset that will be used as a input to our models, and provide us better results.

For embedding, we tried Word2Vec, TFIDF and GloVe. We could see that Glove with 300D embedding gave the best results in model evaluation, hence chose the same for final preprocessing.

We have tried different models i.e., ML (Supervised classification models such as Naive Bayes, Random Forest, XGB, SVM) and NLP based models (BERT).

3.Step-by-step walk through the solution.

Describe the steps you took to solve the problem. What did you find at each stage, and how did it inform the next steps? Build up to the final solution.

Walkthrough :

Firstly we have analyzed the data by importing it then label encoding the columns like industry sector, gender, critical risk etc to get a better understanding of the data and get a visualization of the variance in the data.. Our target variable is 'Custom Accident Level', which we have made by combing the two features 'Accident Level' and 'Potential Accident Level'. We have removed Date column which was unnecessary in the dataset a it served no purpose for our analysis.

And then we found out that there was high imbalance in the data, so we have performed data augmentation and we have used various augmentation techniques like Contextual Word Embedding

Augmentation, Random Word Insertion and Synonym Augmentation. Inside these augmentation techniques also we have used various sub-techniques like Insertion, Deletion, Substitution and Replacement in order to avoid redundancy, and to produce a balanced and augmented dataset that will be used as input to our models, and provide us better results.

Since the aim is to build NLP based chatbot as described in the report, we drop all columns except the **Description** and the **Target Variable (Accident Level).**

We had two data to feed to our different models.

1. **x_train_ml** and **y_train_ml** can be used in **Machine Learning Models**

2. **dataset_chatbot** can be used in **RNN/LSTM Classifier. dataset_chatbot** has its **Stop Words Removed**, **Lemmatized Contextual Word Embedding Augmentation (ContextualWordEmbsAug)** is used in **dataset_chatbot** for handling **Imbalanced Dataset**

Then we fed this augmented data to our models. We have tried different models i.e., ML (Supervised classification models such as Naive Bayes, Random Forest, XGB, SVM) and NLP based models (BERT).

We got our best results from Random Forest, which is a decision tree based ensemble machine learning model that uses ensembling. In prediction problem involving unstructured data(images, text, etc) artificial neural networks tends to outperform all other algorithms, hence we tried out BERT. We have pickled Random Forest model for our chatbot.

4) Model evaluation:

Describe the final model in detail. What was the objective, what parameters were prominent, and how did you evaluate the success of your models?

As per all the metrics (which include accuracy, precision, and recall), the model that performed best was DistilBert.

But since it could not be pickled or saved, and running the model even once consumes a lot of time, we could not rework on it, hence finalized Random Forest Classifier.

This model was performing best among other Machine Learning models.

For each model, hyperparameter tuning was performed before final comparison of different models.

Finally Random Forest Classifier was compared with other models post cross validations.

Even after hyperparameter tuning, Random Forest had a significant difference in accuracy compared to other models.

Other metrics from in the Classification report also favored Random Forest.

Unfortunately, all the comparisons could not be integrated in the final report.

5) Comparison to benchmark:

How does your final solution compare to the benchmark you laid out at the outset? Did you improve on the benchmark? Why or why not?

The final model is the same benchmark model since it was chosen after comparing with other models, only after tuning and cross validations instead of assuming the best model and further improving it.

Hence there can't be further improvements performed on it.

As mentioned in point #4, there is a significant difference in the final model, with a minimum of 10% better performance compared to the next best model.

6) Visualizations:

In addition to quantifying your model and the solution, please include all relevant visualizations that support the ideas/insights that you gleaned from the data.

The predictor variable was visualized along with the target variable (accident level) through histplot to visualize the distribution of data across different accident levels.

Through this graph we could easily spot the huge imbalance, hence augmented the train split of the predictor variable to get a balanced dataset for training.

Visualizations for the predictor variable separately were done using Word Cloud, before data augmentation was done.

It was done for whole the predictor dataset and also separately for each corresponding Accident Level.

Through this we could see that the words were little generic but related to operations in industries, in the lower accident levels.

One peculiar observation is that the words 'left' and 'hand' were hugely repetitive in most distributions, and the word 'right' also was recurrent.

As we skim through higher accident levels, the verbiage was more and more specific to the tools used, and changed as per the levels.

7. Implications
How does your solution affect the problem in the domain or business? What recommendations would you make, and with what level of confidence?

Since it's not explicitly mentioned about where we will use the chatbot. We assumed that this will be deployed on a platform where industrial accidents are reported quite often and this chatbot application will be used to analyse the problem statement and predict the intensity of the problem/accident using various Natural Language Processing algorithms, Machine Learning Algorithms. In the entire process we have tested various users/industries problem statements and have trained the model on the same therefore with a very high level of confidence we can say that our model can perform very well in assessing the level of danger involved in industries.

8. Limitations
What are the limitations of your solution? Where does your model fall short in the real world? What can you do to enhance the solution?

The main limitation of the modem is the practical use case. Since we are dealing with industrial accidents which is a typical emergency, it is not practical to implement a chatbot with no human intervention to take care of such emergencies. Also, the is chatbot cannot be a full end to end solution for problems but these can form a very essential part of a larger project.

9. Closing Reflections
What have you learned from the process? What you do differently next time?

We as a team have learnt how to plan the process from start to end and forecast our availability. Collaborating our codes and co-ordinating with each other and setting major milestones for the entire project. Next time I will allocate a significant time for the planning and the process of improving the model performance rather than trying to code the logic. Overall, we learnt a lot of minute things with respect to programming, logic, managing computational requirements, resources, etc and we are sure we will able to perform better in model building the next time.