**INTERIM REPORT:**

1) Problem Statement, Data and Findings:

We have a dataset from the databases of the biggest industries in Brazil and in other parts of the world. We have an urgent need to help companies understand why there are so many accidents happening in industrial plants in spite of the technical advancements we have in today's world.

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

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.

2) Preprocessing: 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.

3) Deciding Models and Model Building

Based on the nature of the problem, decide what algorithms will be suitable and why?

Experiment with different algorithms and get the performance of each algorithm.

We will try different models i.e., ML (Supervised classification models such as Naive Bayes, Random Forest, etc.) and NLP based models (BERT, LSTM , Transfer learning model) with different embeddings techniques like Word2Vec and TFIDF; but for now we have implemented Naive Bayes and random Forest post GloVe embedding.

Random Forests is attempted since decision tree performs best on imbalanced data. They work by learning a hierarchy of if/else questions and this can force both classes to be addressed.

LSTM stands for Long-Short Term Memory. LSTM is a type of recurrent neural network but is better than traditional recurrent neural networks in terms of memory. Having a good hold over memorizing certain patterns LSTMs perform fairly better. LSTMs efficiently improves performance by memorizing the relevant information that is important and finds the pattern.

And in our case Random Forest gave train accuracy close to 1

4. How to improve your model performance?

What are the approaches you can take to improve your model? Can you do some feature

selection, data manipulation and model improvements.

To Improve our model performance: -

1. We could use Model with Hyperparameter Tuning with different values for the parameters’ splits, random state, classifier name, etc.

2. We could use different sampling techniques with our models to test their performance, with and without sampling

3. Convert Classification to Numeric problem:

We can create a classification model that uses categorical columns and tf-idf features from accident description and label encoded target variable. We can use simple densely connected neural networks to make predictions.

Improvements needed in dataset:

1. The Dataset is relatively small for model building.

2. Description column is not structured; some words could be used to describe accurate description.

3. The provided Dataset is unbalanced.

4. The severity of the accident is not evident from the name of body parts or the accident name, there should be another variable which describes the level of the damage or the severity.