**AIML**

**CAPSTONE PROJECT - INTERIM REPORT**

**NLP CHATBOT For Industry Safety**

**Team Mentor**

Rohit Gupta

**Team Members**

Sarita Ghadge

Vikram Rai

Rohan Bhattacharya

Meetu Chandra

# Problem Statement

Major Industries in Brazil is facing multiple workplace hazards leading to multiple accidents, some even leading to death. The aim of this project is to build an AI powered Chatbot for employees and stakeholders to understand the various reason for the accidents. This can help the employees take the necessary precautions and for the stakeholders to identify the root cause for the accidents and put a fix in place.

The Chatbot will be powered by an AI model that has been trained to classify the accident, give a summary of the various scenarios when the accident happened in past.

# Approach

An NLP based AI model will be trained on industry data from **3 countries** and **12 plants.** The data is from industries operating majorily in Metal and mining. Machine learning and deep learning models will be developed and compared with cleaned data.

The below models will be used and compared.

**SVM classifier** - Useful for tasks with high-dimensional data and clear margins between classes.

**Random Forests** or **XGBoost** - Work well for structured data representations like TF-IDF or word embeddings.

**Recurrent Neural Networks (RNNs)/LSTM** - Good for sequential data processing.

**Transformers (e.g., BERT, DistilBERT)** - State-of-the-art for modern NLP classification tasks.

# Data Overview

# We analysed data from one of the largest industries in Brazil, which consists of 425 accident records from January 2016 to September 2017 across 3 countries and 12 locations. The dataset includes key information provided in the table below

|  |  |  |
| --- | --- | --- |
| **Column** | **Description about the column** | **Observations** |
| Data | timestamp or time/date information |  |
| Countries | which country the accident occurred (anonymised) |  |
| Local | the city where the manufacturing plant is located (anonymised) |  |
| Industry sector | which sector the plant belongs to |  |
| Accident level | from I to VI, it registers how severe was the accident (I means not severe but VI means very severe) |  |
| Potential Accident Level | Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors involved in the accident) |  |
| Genre | if the person is male of female |  |
| Employee or Third Party | if the injured person is an employee or a third party |  |
| Critical Risk | some description of the risk involved in the accident |  |
| Description | Detailed description of how the accident happened. |  |
| Data | timestamp or time/date information |  |

# Solution

1. Summary of the Approach to EDA and Pre-processing Include any insightful visualization you have teased out of the data. If you’ve identified particularly meaningful features, interactions or summary data, share them and explain what you noticed. Visual displays are powerful when used well, so think carefully about what information the display conveys.
2. 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.
3. How to improve your model performance? meetuchandra@gmail.com 2VA1NF695Q What are the approaches you can take to improve your model? Can you do some feature selection, data manipulation and model improvements. Provide your code and as much as visualizations you can share to describe what you have done so far. This file is meant for personal use by meetuchandra@gmail.com only. Sharing or publishing the contents in part or full is liable for legal action.