CHATBOT INTERFACE USING NLP

POST GRADUATE PROGRAM IN

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

TEXAS McCombs The University of Texas at Austin McCombs School of Business | GREAT LAKES EXECUTIVE LEARNING | Great Learning POWER AHEAD

Dec 2021

Group Capstone Project for PGP AIML

Project Team JAN A G:6 (NLP-2)

Mohana Krishna Suryadevara, Neethu Jacob, Premkumar Coimbatore Govindan, Rakesh Kumar, Varun Prakash

Chatbot Interface using NLP

GROUP CAPSTONE PROJECT FOR PGP AIML

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 What is a Chatbot?

A Chatbot is an Artificial Intelligence (AI) software that can simulate a conversation (or a chat) with a User in natural language through messaging applications, websites, mobile apps, or through the telephone. A Chatbot is often described as one of the most advanced and promising expressions of interaction between humans and machines.

From a technological point of view, a Chatbot only represents the natural evolution of a question & answer system leveraging Natural Language Processing (NLP). Formulating responses to questions in natural language is one of the most typical use cases of NLP applied in various enterprises' end-user applications.

1.2 How does a Chatbot work?

There are two different tasks at the core of a Chatbot:

- User Request Analysis – Chatbot analyses the User's request to identify the user intent and to extract relevant entities
- Returning the Response – Chatbot selects and provides the most appropriate response back to the User – a generic pre-defined answer, a clarifying question to seek further information from the User, or more advanced capabilities that involve understanding the context, leveraging the knowledge base of responses, performing some action etc.

There are different approaches and tools to develop a Chatbot. Depending on the use case, some Chatbot technologies are more appropriate than others. To create an effective Chatbot, a combination of different AI techniques such as Natural Language Processing (NLP), Machine Learning (ML), and Semantic Understanding may be the viable option.

1.3 Types of Chatbots

At the highest level, there are three types of Chatbots that Consumers' experience.

1.3.1 Rules-based Chatbot

A Rule-based Chatbot follows pre-designed rules, often built using a graphical user interface (GUI) where a bot builder designs paths using a decision tree.

1.3.2 AI Chatbots

AI Chatbots automatically learn and adapt their responses after an initial training period by a bot developer.

1.3.3 Live Chat

Live Chat is primarily used by Sales & Sales Development teams. It can also be used by Customer Support organizations, as Live Chat is a more simplistic chat option to answer questions in real-time by a human agent.

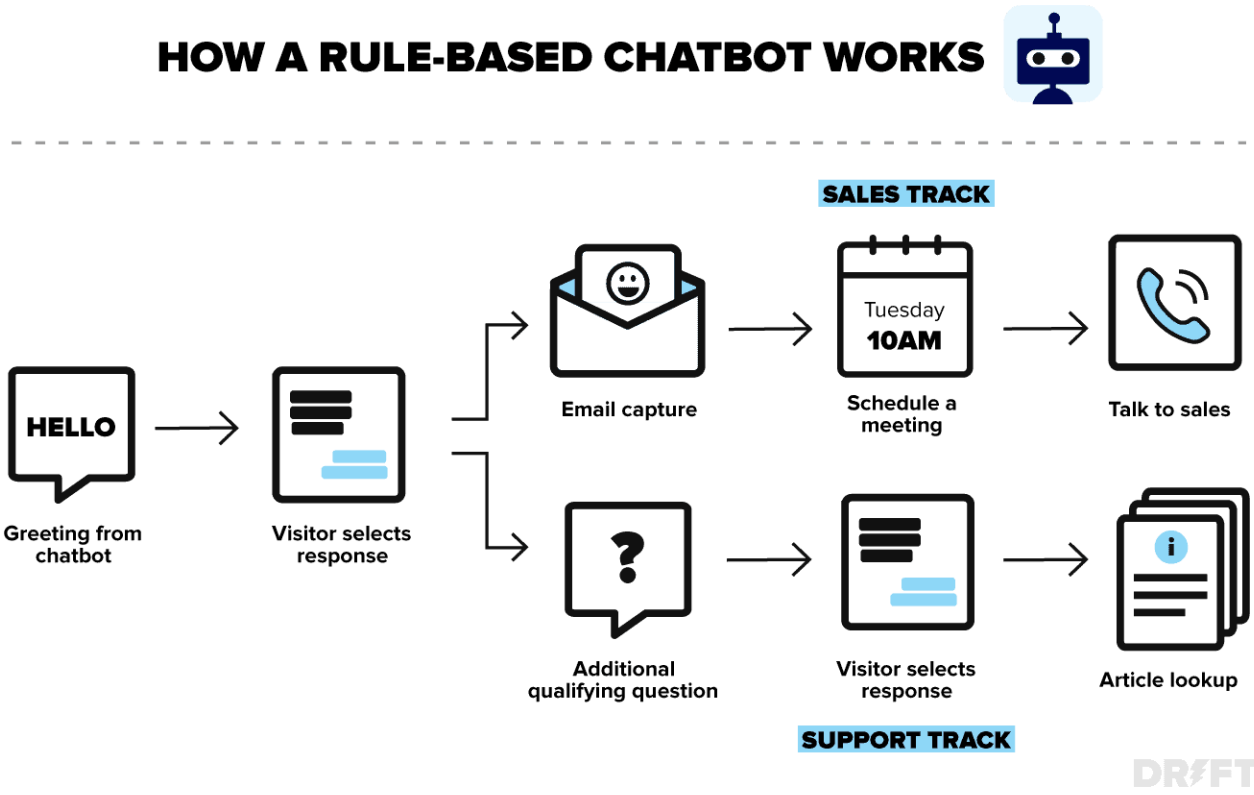

Figure 1: How a Rule-Based Chatbot Works

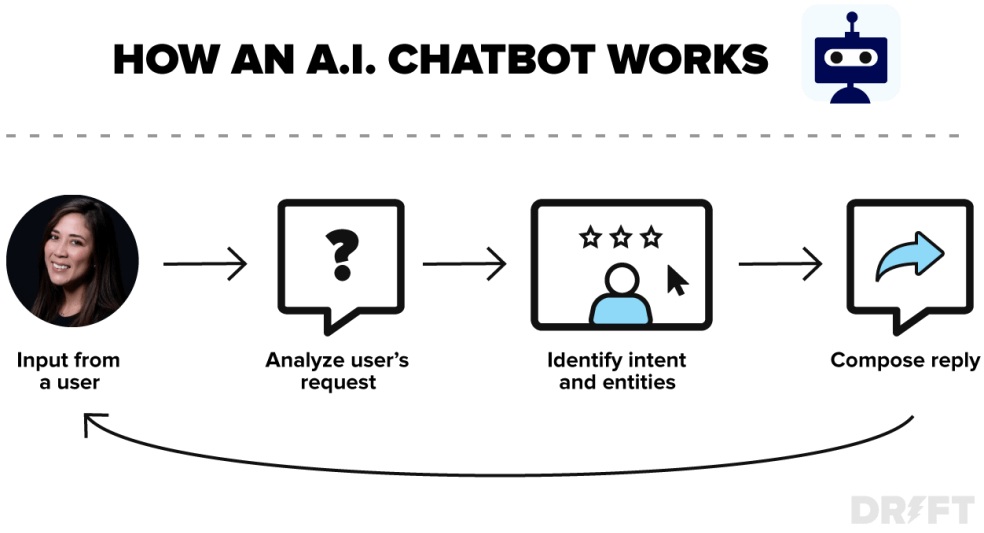

Figure 2: How an AI Chatbot Works

1.4 Why is a Chatbot important?

Chatbot applications streamline interactions between people and services, enhancing customer experience. At the same time, they offer companies new opportunities to improve the customers engagement process and operational efficiency by reducing the typical cost of customer service.

To be successful, a Chatbot solution should be able to effectively perform both of these tasks. Human support plays a key role here: regardless of the kind of approach and the platform, human intervention is crucial in configuring, training and optimizing the Chatbot system.

Chatbots have become common and standard in several industry segments and sectors, with tremendous advancements in cloud computing infrastructure, higher speed and bandwidth in internet streaming, increased penetration of smartphones and evolution of plethora of smart web-based and mobile applications in a mobile-first world.

1.5 Evolution of Chatbots

ELIZA is an early natural language processing computer program created from 1964 to 1966 at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum. Eliza simulated conversation by using a "pattern matching" and substitution methodology that gave users an illusion of understanding on the part of the program, but had no built-in framework for contextualizing events.

Gartner's take is that Conversational AI platforms are the foundational technology for development of Chatbots and Virtual Assistants (VA) that are applications on a sophistication continuum.

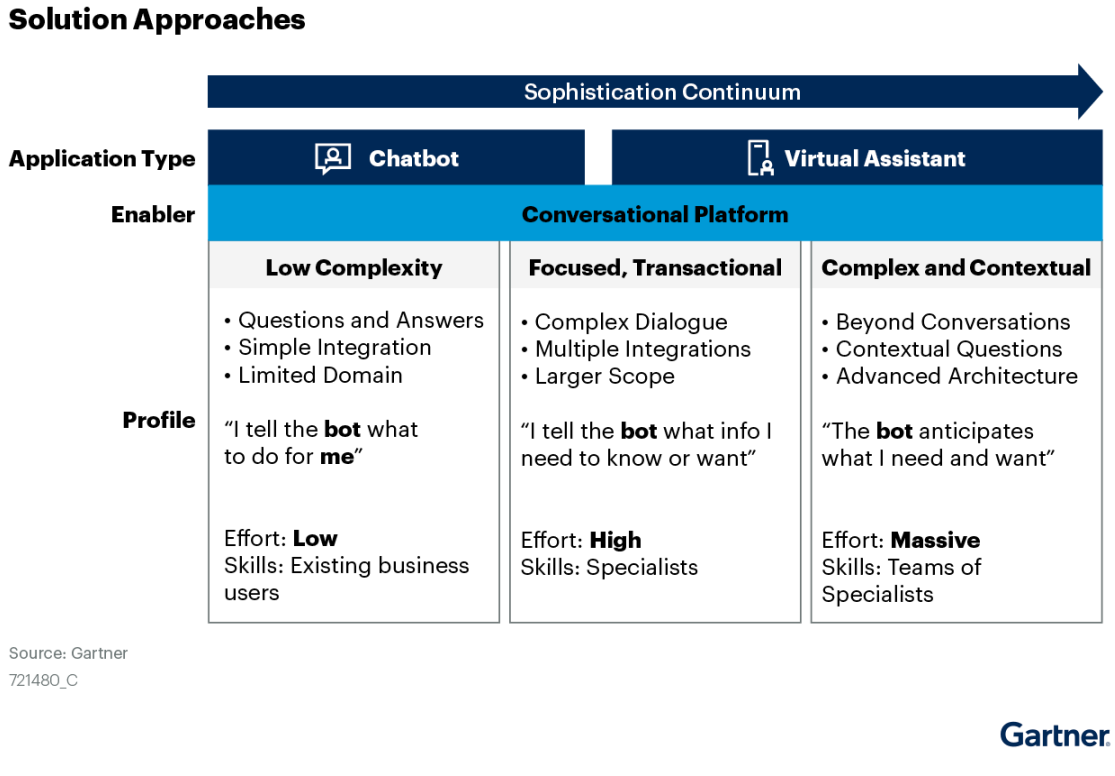

Figure 3: Gartner - Sophistication Continuum of Chatbot and Virtual Assistant

The NLP Pipeline schematic from Gartner illustrates the role of NLP in enabling Chatbots on Intent Prediction and Entity Extraction leading to Clarifying Questions or Responses to the Utterances from User.

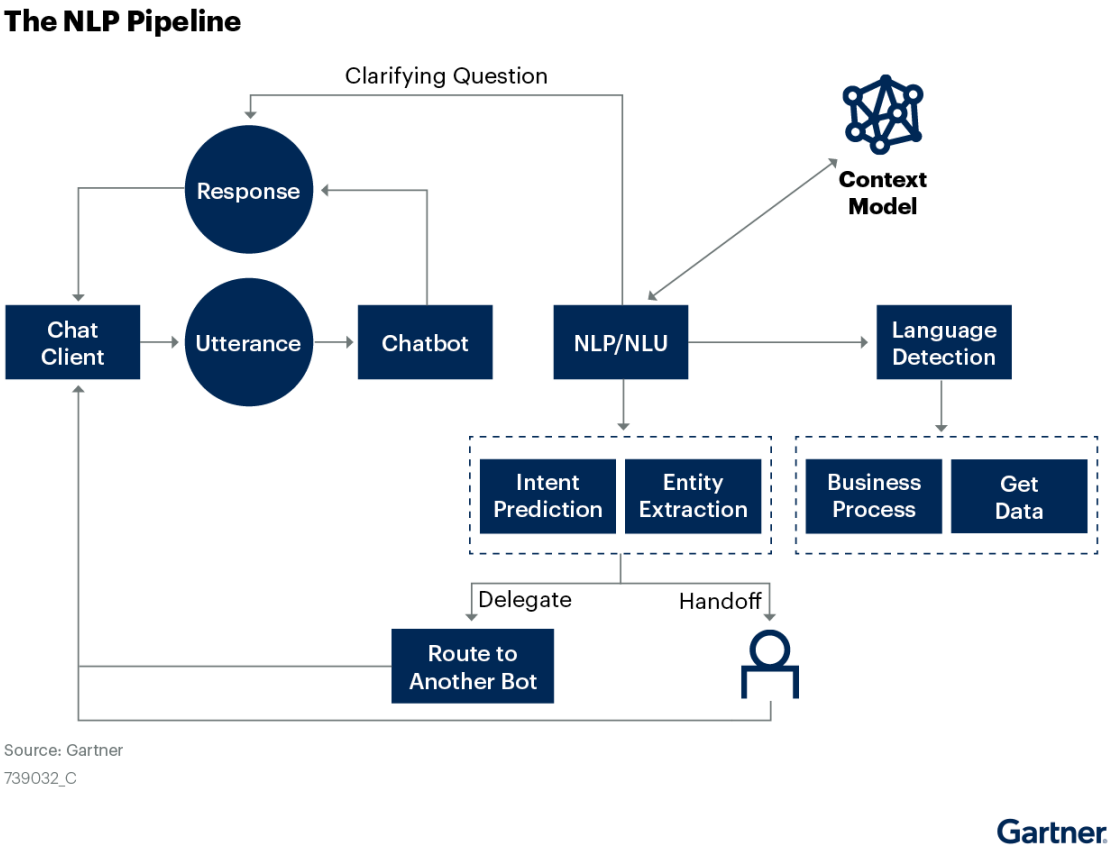

Figure 4: Gartner - NLP Pipeline

1.6 Chatbot Capabilities to Consider

Chatbots can range from fairly simple to highly complex in terms of their architecture and capabilities. In this section, we explore some of the advanced features you can expect in enterprise-grade Chatbots that deliver outstanding customer experience while lowering costs and effort through Artificial Intelligence and automation.

1.6.1 Interactions

First, you need to master the art of conversation. You must customize your Chatbot to be a conversationalist - neither dull nor pervasive. When you create your Chatbot the right way, on the right platform, and with the right Chatbot script, you'll discover how amazing bots are as your Customers' Personal Assistants.

1.6.2 Conversational Maturity

The greatest advantage of building a customizable Chatbot is that you can train it to be how your Customers want it. Your Chatbots can collect User data for you to analyze the average maturity level of your audience and maintain consistency within acceptable deviations. As your Chatbots gain insights, you'll understand how to edit your Chatbot to suit your audience.

1.6.3 Emotional Intelligence

Emotions are part and parcel of life. As much as we try and set them apart, they will be in there, somewhere. Since Chatbots are the primary interface between your Business and Customers, try and build a relationship between your Chatbots and your Customers. With Sentiment Analysis, Chatbots can pick up the underlying emotions and respond in an appropriate manner.

1.6.4 Trainable Intelligence

This is among the most exciting Chatbot features. A Chatbot must be able to perform complex reasoning on its own, without human interference. Instead of manually adding and updating FAQs, you can simply load your knowledge base to the Chatbot. The Chatbot parses through the information and can provide a suitable answer within seconds. And as we all know, the faster and more appropriate the response, the more loyal the Customer.

1.6.5 Easy Omnichannel Deployment

WhatsApp, Facebook, Instagram, Twitter, Telegram – social media platforms such as these have become the means of corporate survival. To be absent on social media is like turning your back on 70% of your customers. Therefore, you must deploy your Chatbot across channels, in addition to your websites, with a consistent and engaging customer experience.

1.6.6 Extensible Integrations

Integrate the Chatbot with your preferred 3rd-party applications like Salesforce, Zendesk, Google Sheets, and more. Generate leads, collect data and achieve maximum Chatbot functionality, with support for native integrations and custom integrations.

1.6.7 Rich Contextual History

Leverage state-of-the-art NLP Engines that use previous conversations to make future conversations better.

1.6.8 Training Made Easy

Building a chatbot is never a one-stop process. While good innovators create good products, great innovators continuously test their products. Ensure that the bot that you build is well-tested and framed for a seamless customer experience.

1.6.9 Easy Human-takeover

Your chatbot can handle around 80% of your customer queries without human intervention. But, it should be easy for your live agents to take over the more complex conversations.

1.6.10 Robust API

A robust Chatbot API will ensure that your Customers have the freedom and the authority to browse through it and find what they are looking for. It will keep the engagement intact and create an unprecedented experience.

In addition, a Chatbot can also integrate with and invoke external APIs to extend the knowledge base and respond to User requests beyond what it is trained to do. API integration makes it easier for Chatbots to fetch resources and insights from other applications, both inside and outside your organization. In order to extract value from different systems or applications, it is essential to connect your Chatbot with the relevant API.

2 CAPSTONE PROJECT

2.1 Project Overview

2.1.1 Opportunity

The safety of people who work in industries involving operating heavy machinery has always been important. While there has been an evolution in the industry with increased awareness on safety measures, advancement in machinery including automation, as well as improvement in protective gear, accidents do tend to happen due to various avoidable and unavoidable reasons and circumstances.

One of the biggest industries in Brazil and in the world has been collecting data on industrial accidents involving their personnel, that have happened over the years in mining, metal, and other sectors. Based on the description of the event and the critical risk factors identified, the accidents have been classified on severity level. It is an urgent need for industries/companies around the globe to understand why employees still suffer some injuries/accidents in industrial plants. Sometimes they also die in such environment.

With a vast amount of historical data on accidents, leveraging Artificial Intelligence and Machine Learning to create systems that can automate the classification and possible prevention of accidents through actionable data-driven insights is a compelling use case. In addition to the effort, time, and cost savings that can be achieved, the primary purpose of making the workplace safer is driving a lot of investment in AIML based systems to drive business decisions and actions.

2.1.2 Objective

Design a ML/DL based Chatbot utility which can help the professional highlight the safety risk as per the incident description.

The target Users of this Chatbot utility are the Leaders, Supervisors, Workers in the industry who can either understand the accident level classification of an accident that has occurred, or estimate the potential risks involved in a future planned activity.

The target Users can also get some insights on the patterns and trends in accidents to anticipate future occurrences and take some proactive steps to enhance the safety measures, through appropriate safety equipment, training to the personnel, evolving process checklists that can be incorporated in the workplace and the safety drill routines that can prevent avoidable accidents.

2.1.3 Overview of Project Milestones

Table 1: Capstone Project Milestones

Capstone Schedule				
Week	**Date (Fri)**	**Mile Stones**	**Session date (Sat/ Sun)**	**Session Plan**
Week-0	12/11/2021	**Start of capstone.** Release the Capstone problem statement & Grouping	13/14 NOV	No session

Capstone Schedule				
Week	**Date (Fri)**	**Mile Stones**	**Session date (Sat/ Sun)**	**Session Plan**
Week-1	19/11/2021	**Release Interim Report** Group discussion on the respective problem statement released	20/21 NOV	Discussion of Problem Statement, Delegation of responsibilities among team members, Planning the progress and approach to capstone
Week-2	26/11/2021	**Milestone-1:** EDA and Pre-processing (Feature engineering and selection	27/28 NOV	Review progress of team & individuals' contributions, pointing out errors, suggesting alternative methods, clarification of conceptual roadblocks if any.
Week-3	3/12/2021	**Milestone-1:** Modelling (Focus on accuracy and generalization) & interim Report Submission	4/5 DEC	Validate the interim submission and suggest improvements if any. Raise alarms for proactive correction
Week-4	10/12/2021	**Interim report submission (12 DEC)** **Milestone-2:** Evaluation of your model (Comparison of different models, performance tuning) **Release Final Report (10 DEC)**	11/12 DEC	Review the progress towards final reporting
Week-5	17/12/2021	**Milestone-2:** Final model tuning and documentation	18/19 DEC	Validation of the final report and suggestions of correction before submission
Week-6	24/12/2021	**Milestone-2:** Presentation to the mentor	25/26 DEC	Presentation of capstone project by students to mentor
Week-7	31/12/2021	**Final Report submission on Olympus (2ND JAN)**	1/2 JAN	Course Closure – Non-Mentorship session with program Manager

2.1.4 Acknowledgments

As a Project Team, we thank the supportive and proactive engagement from the Program Manager – Shabana Inayeth Khan, who has been diligent in her communication, follow-ups, responding to queries and support requests, setting and clarifying expectations. Shabana also was instrumental in assembling our project team where we each complemented each other on the skill sets and were able to become a project team with the right guidance.

Through the PGP AIML course, and particularly during the Capstone Project, our Project Mentor – Aditya Bandaru – was providing insights, functional and technical guidance, and challenging us as a team to think big picture, while exploring the architectural and design choices to evolve and implement an optimal solution. Aditya took a very pragmatic approach drawing on his experience in the industry and pointed us towards enterprise-grade platforms and packages that are leveraged for AIML solutions, while focusing on the basics of Exploratory Data Analysis, Hypothesis Testing, Feature Engineering, Model Selection and Tuning, User Interface and Interactions, Deployment Architecture, Testing various business and technical scenarios. Aditya's ability to see a piece of code for the first time and help troubleshoot it to improve both the quality and performance was definitely a blessing for our Project Team.

In addition, the various faculty and guest speakers who shared their knowledge in industry sessions helped us relate and apply what we are learning to real-world use cases.

2.2 Project Team

2.2.1 Mohana Krishna Suryadevara
2.2.2 Neethu Jacob
2.2.3 PremKumar Coimbatore Govindan

I had majored in Mathematics and did my masters in Computer Science back in 2003. Over the past two decades, technology has grown tremendously in leaps and bounds, particularly in Data Science, Artificial Intelligence and Machine Learning. As a Business-Technology Leader looking to grow my career in the trending technology area of Data Science and Analytics, I signed up for the PGP-AIML course from Great Learning to enhance my knowledge and understanding of the concepts, techniques, challenges, and platforms to ideate and implement future-state products and solutions that disrupt the industry and deliver value to the consumers, customers, company and investors.

I chose the NLP Chatbot project as I am particularly interested in the human interaction with machine through conversational AI which is a trending topic with compelling use cases and wide spread adoption in B2C or a B2B context. Beyond the Chatbot application, the larger scope of NLP is something that excites me, what with language being the oldest and the most important invention of humankind, particularly with so many popular languages across the world, and a need for interpreting and translating languages in a global economy has tremendous potential.

2.2.4 Rakesh Kumar
2.2.5 Varun Prakash

3 ANALYSIS

3.1 Exploratory Data Analysis (EDA) & Visualization

3.1.1 Data Collection

The input was a CSV file (Comma-separated Values) with the first row containing the Column Names. Using the read_csv method from Pandas library, imported the data into a Pandas Dataframe. The Dataframe had 425 rows and 11 columns on successful data import without any errors. Initial observations on the structure and content of the dataset are described in the table below.

Table 2: Brazil Industrial Safety Dataset with Accident Descriptions

Column Name	Column Description	Column Data Type	Initial Observations
Unnamed: 0	Unknown	Int64 (Number)	Row Number (0,1, …,424)
Data	timestamp or time/date information	Object (String)	Date in YYYY-MM-DD format with timestamp as 00:00:00
Countries	which country the accident occurred (anonymized)	Object (String)	Categorical values of 3 different Countries
Local	the city where the manufacturing plant is located (anonymized)	Object (String)	Categorical values of 12 different Cities (across the 3 Countries)
Industry Sector	which sector the plant belongs to	Object (String)	Categorical values of 3 different industry sectors

Column Name	Column Description	Column Data Type	Initial Observations
Accident Level	from I to VI, it registers how severe was the accident (I means not severe but VI means very severe)	Object (String)	Categorical values of 5 different levels found in Dataset (of the 6 possible values mentioned in column description)
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)	Object (String)	Categorical values of 6 different levels found in Dataset (as per the values mentioned in column description)
Genre	if the person is Male or Female	Object (String)	Categorical values of 2 different Genders (Male, Female)
Employee or Third Party	if the injured person is an employee or a third party	Object (String)	Categorical values of 3 different employee types
Critical Risk	some description of the risk involved in the accident	Object (String)	Categorial values of more than 30 different critical risk descriptions
Description	Detailed description of how the accident happened	Object (String)	Descriptive text with several sentences.

3.1.2 Data Cleanup and Pre-Processing

3.1.2.1 Dropping Columns

The first column "Unnamed: 0" was dropped as it was a sequence indicating the row number which is not useful for analysis.

3.1.2.2 Renaming Columns

Some of the column names were vague or incorrect when compared with the specification in the problem statement. Renamed columns to a more meaningful name that captures the nature of the values more accurately and helps use the right labels for data visualization.

Table 3: Renaming Columns

Old Column Name	New Column Name
Data	AccidentDate
Countries	Country
Industry Sector	IndustrySector
Accident Level	AccidentLevel
Potential Accident Level	PotentialAccidentLevel
Genre	Gender
Employee or Third Party	EmployeeType
Critical Risk	CriticalRisk

3.1.2.3 Duplicate Check

There were 7 duplicate rows of data in the input. The duplicates were removed from the dataset. After dropping a column and removing the duplicate rows of data, the Dataframe had 418 rows and 10 columns.

3.1.2.4 Null Values Check

There were no NULL values at all in the dataset. All columns had values for all the rows.

3.1.2.5 Check for Outliers

There were no outliers in the data. Thus, there was no need for handling outliers.

- The AccidentDate column had date values in the years of 2016 (all months) and 2017 (until July).
- The Description was a free text field
- All other columns were categorical in nature. The CriticalRisk field could have been better captured. Most values were "Others" which wasn't useful.

3.1.2.6 Data Types

AccidentDate column was converted to a Date datatype while ignoring the time information, which was always 00:00:00

3.1.2.7 Computed Columns

In order to facilitate exploratory data analysis, hypothesis testing, and feature engineering for model evaluation and selection, following computed columns were added to the Dataframe:

Table 4: Computed Columns added to Dataframe

Computed Column Name	Computation Logic	Description
AccidentYear	Year value from AccidentDate	Year when the Accident occurred in YYYY format
AccidentMonth	Month value from AccidentDate	Month when the Accident occurred in MM format
AccidentDay	Day of the Month value from AccidentDate	Day of the Month when the Accident occurred in DD format
AccidentDayOfWeek	Day of the Week value from AccidentDate	Day of the Week when the Accident occurred (Sunday, Monday, …)
AccidentWeekOfYear	Week Number of the Year from AccidentDate	Week Number of the Year when the Accident occurred – a numeric value in the range of (1, 2, …, 53)
Season	Computed based on the Month of the Year and seasons of Brazil. • 9, 10, 11 => Spring • 12, 1, 2 => Summer • 3, 4, 5 => Autumn • 6, 7, 8 => Winter	The season of the Year. The intent was to understand if the season, and consequently the weather patterns, have any impact on the frequency and severity of Accidents.
IsHoliday	Computed by looking up the List of National Holidays in Brazil	Whether the Date of Accident was a Holiday or not.

3.1.3 Variable Identification

- Target Variable(s): Combine AccidentLevel and PotentialAccidentLevel to create a new Target Variable (AccidentLevel)
- Input Variable(s): AccidentDate, Country, Local, IndustrySector, Gender, EmployeType, CriticalRisk, Description

3.1.4 Univariate Analysis

3.1.4.1 Country

Table 5: Accidents reported by Country

Country	Count	%
Country_01	248	59%
Country_02	129	31%
Country_03	41	10%

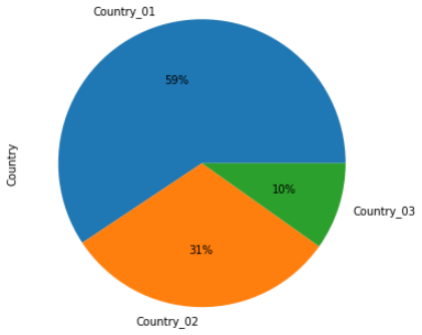

Figure 5: Accidents reported by Country

Country_01 has the most instances of Accidents reported. Country_03 has the least instances of Accidents reported.

3.1.4.2 Local

Table 6: Accidents reported by Local

Local	Count	%	Cumulative %
Local_03	89	21.3%	21.3%
Local_05	59	14.1%	35.4%
Local_01	56	13.4%	48.8%
Local_04	55	13.2%	62.0%
Local_06	46	11.0%	73.0%
Local_10	41	9.8%	82.8%
Local_08	27	6.5%	89.2%
Local_02	23	5.5%	94.7%
Local_07	14	3.3%	98.1%
Local_12	4	1.0%	99.0%
Local_09	2	0.5%	99.5%
Local_11	2	0.5%	100.0%

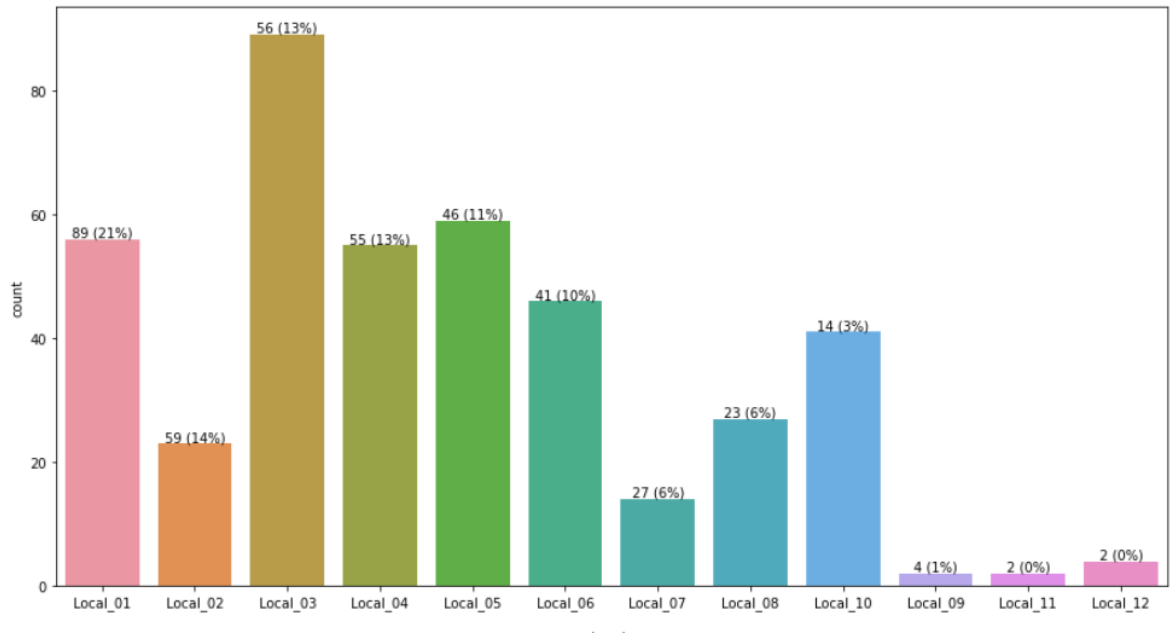

Figure 6: Accidents reported by Local

Local_03 has the most instances of Accidents reported, while Local_09 and Local_11 have the least instances of Accidents reported.

The Top 3 Locals (25% of all Locals) accounted for approximately 50% of Accidents reported. The Top 6 Locals (50% of all Locals) accounted for over 80% of Accidents reported.

3.1.4.3 Gender

Table 7: Accidents reported based on Gender of Person injured

Gender	Count	%
Male	396	94.7%
Female	22	5.3%

The data is highly skewed on the Gender values. Most of the Accidents reported (~95%) involve Males. Since the Gender ratio of the Total Workforce is unknown, there can be no meaningful conclusions drawn on whether Gender plays a role in the proportion of Accidents being reported.

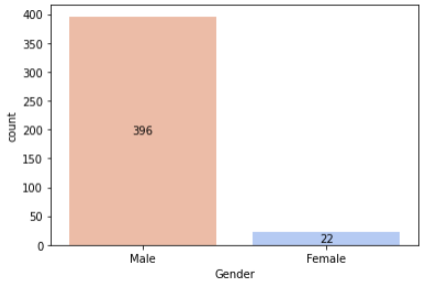

Figure 7: Accidents reported by Gender of Person injured

3.1.4.4 Industry Sector

Table 8: Accidents reported by Industry Sector

Industry Sector	Count	%
Mining	237	57.0%
Metals	134	32.0%
Others	47	11.0%

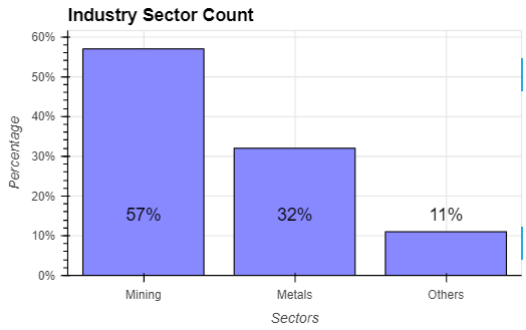

Figure 8: Accidents reported by Industry Sector

Mining sector contributes to highest number of Accidents reported, followed by Metals sector, which together make up 89% of all Industrial accidents reported in the Dataset.

3.1.4.5 Safety Risk Classification

The Dataset revolves around the Industrial Safety Risk Classification expressed in terms of two Target Variables, namely, "Accident Level" and "Potential Accident Level" with values ranging from I (lowest risk) to VI (highest risk).

In the Dataset, the general pattern was that the value for Potential Accident Level was typically greater than the value for Accident Level. Only 1 Accident was classified as having the highest level of VI on the Potential Accident Level, while 0 Accidents were reported with having a level of VI on Accident Level.

Table 9: Accidents reported by Accident Level (Lowest to Highest)

Accident Level (Lowest to Highest)	Count	%
Accident Level – I	309	74%
Accident Level - II	40	10%
Accident Level - III	31	7%
Accident Level - IV	30	7%
Accident Level - V	8	2%
Accident Level - VI	0	0%

Table 10: Accidents reported by Potential Accident Level (Lowest to Highest)

Potential Accident Level (Lowest to Highest)	Count	%
Potential Accident Level – I	45	11%
Potential Accident Level - II	95	23%
Potential Accident Level - III	106	25%
Potential Accident Level - IV	141	34%
Potential Accident Level - V	30	7%
Potential Accident Level - VI	1	~ 0%

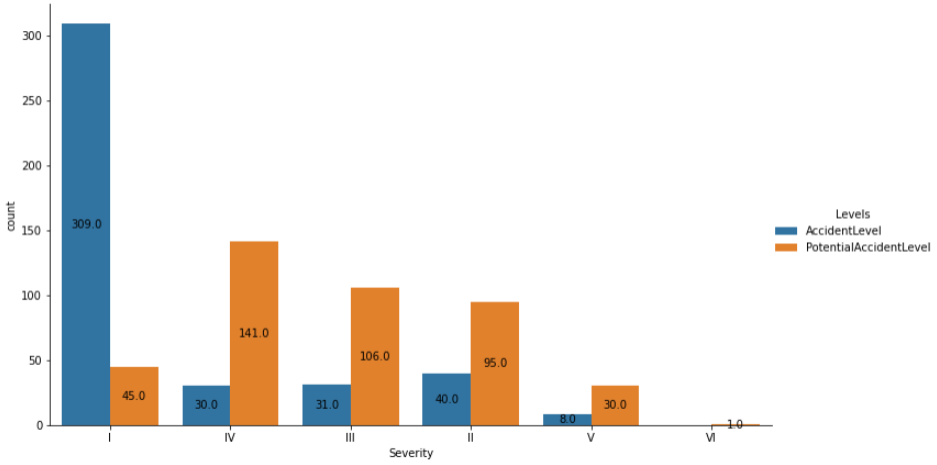

Figure 9: Accidents by Level - Reported and Potential

Our interpretation is that either some precautionary controls, or safety gear, or timely reaction, or fortune would have played a role in toning down the severity of the injury that could have been caused by the accident. However, it could also be a case of people underreporting the severity of the accident to avoid panic or penalties. It is recommended to pay attention to the Potential Accident Level rating of accidents and activities that carry a similar risk profile to prevent or prepare for the worst-case scenario.

Gathering more data over a longer period of time, benchmarking the trend against other Brazilian companies in similar Industry Sectors, or against other Countries in similar Industry Sector can help validate the assumptions and guidelines for the assignment of Accident Level and Potential Accident Level.

3.1.4.6 Employee Type

Table 11: Accidents reported based on Employee Type of Person injured

Employee Type	Count	%
Third Party	185	44%
Employee	178	43%
Third Party (Remote)	55	13%

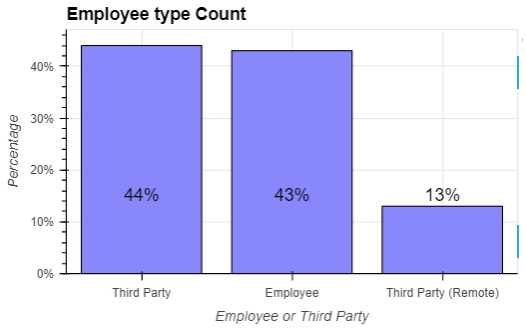

Figure 10: Accidents reported by Employee Type of Person injured

Number of Accidents involving Employee and Third Party are almost on par with each other. Since the overall ratio of number of Employees and Third Party Personnel is unknown, we can't draw any further conclusions. Third Party Personnel who are in remote location are also prone to Accidents to an extent (13% of the cases). It is not known if any of the Employees too are based in remote locations or only work onsite in the plants / company facilities.

3.1.4.7 Critical Risk

Table 12: Accidents reported classified by Critical Risk

Critical Risk	Count	%
Others	229	55%
Pressed	24	6%
Manual Tools	20	5%
Chemical substances	17	4%
Cut	14	3%
Projection	13	3%
Venomous Animals	13	3%
Bees	10	2%
Fall	9	2%
Vehicles and Mobile Equipment	8	2%
Pressurized Systems	7	2%
remains of choco	7	2%
Fall prevention (same level)	7	2%
Suspended Loads	6	1%
Fall prevention	6	1%
Pressurized Systems / Chemical Substances	3	1%
Liquid Metal	3	1%
Blocking and isolation ofenergies	3	1%
Power lock	3	1%
Machine Protection	2	0%
Electrical Shock	2	0%
Poll	1	0%
\nNot applicable	1	0%

Critical Risk	Count	%
Confined space	1	0%
Burn	1	0%
Individual protection equipment	1	0%
Projection/Burning	1	0%
Projection/Choco	1	0%
Plates	1	0%
Traffic	1	0%
Projection of fragments	1	0%
Electrical installation	1	0%
Projection/Manual Tools	1	0%

The data captured in Critical Risk column is skewed with majority of them being classified as “Others”. Moreover, there is vagueness and overlaps in some of the other categories which is probably causing people to choose “Others” more often. There needs to be a better standard for the Critical Risk classification and better discipline in capturing a more meaningful value instead of “Others”.

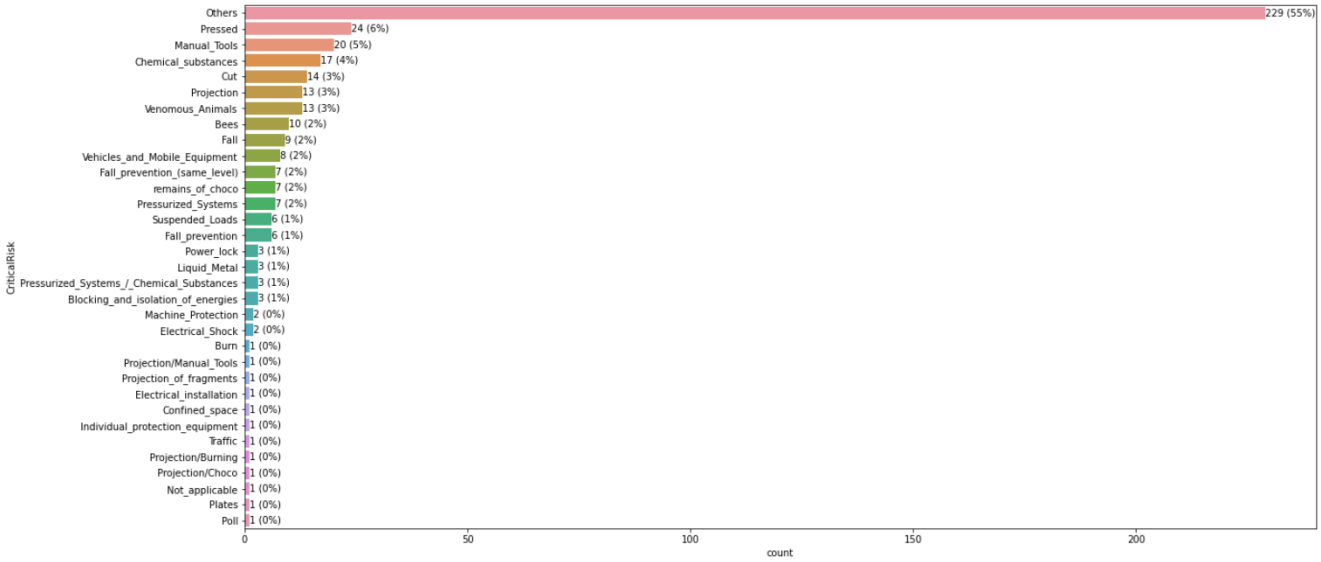

Figure 11: Accidents by Critical Risk classification

3.1.4.8 Accident Date

The Accident Dates range from 1-Jan-2016 to 9-Jul-2017. Year 2016 has accident reported in all the 12 months of the year. Year 2017 has accidents reported only about half of the year. Thus, when we analyze the data at the year level, 2017 has lower counts than 2016.

In Year 2016, a total of 283 Accidents were reported, which is ~68% of all the Accidents recorded in the Dataset. The rest of the Accidents were reported in Year 2017 (only until July) with 135 Accidents that is the balance 32% of the Accidents. Thus, overall volume for full year 2016 is almost on par with volume for the half year 2017.

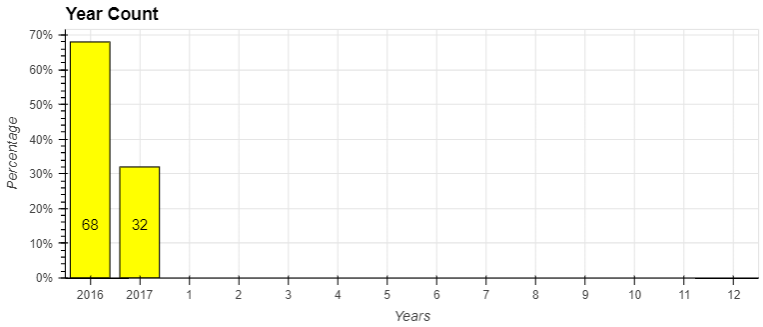

Figure 12: Accidents by Year

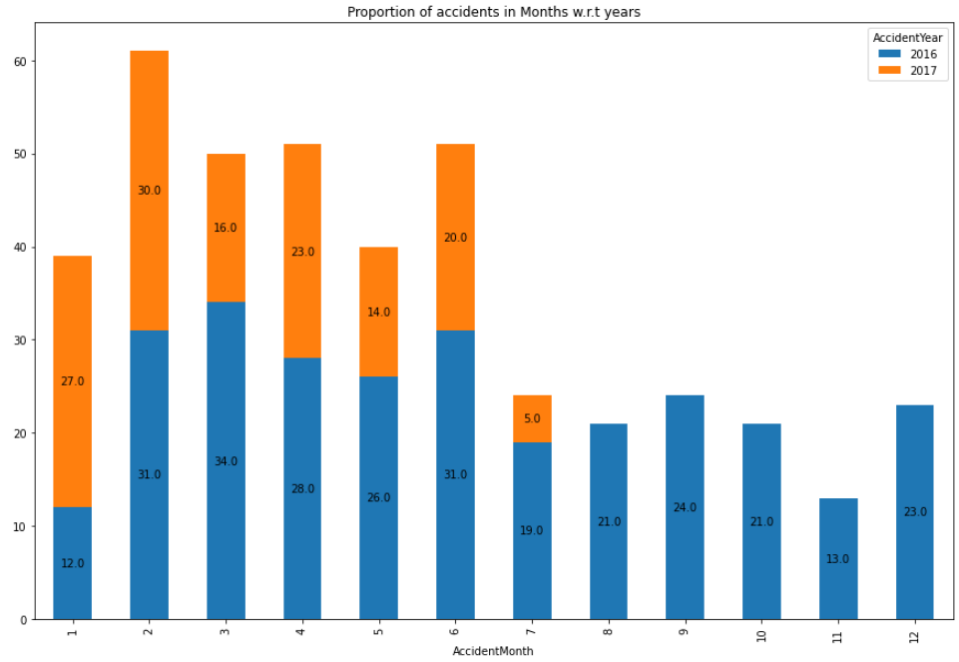

Figure 13: Accidents by Month with Year-wise split

Accidents reported in January 2017 is alarmingly more than twice of what was reported in January 2016. February 2016 and February 2017 has similar high count of ~30. March 2017 is showing excellent improvement as compared to March 2016 with less than half the number of Accidents reported. The Accident trend of April, May, June 2017 is consistently lower that corresponding months in 2016.

Overall, in 2016, Number of Accidents reported in the Months of February (2nd Month) through June (6th Month) is higher than the other months of the Year. Overall, in 2017, among the first 7 months where Accidents were reported, January (1st Month) and February (2nd Month) have a higher proportion of Accidents reported.

Among the Months where Accidents were reported in both years, number of Accidents in 2017 were lesser than those in 2016. Particularly, in the Month of March, there were much lesser Accidents reported in 2017 when compared to 2016.

Across the Months in 2016 and 2017, when we analyze the Number of Accidents reported by Day of the Month, the following observations are evident:

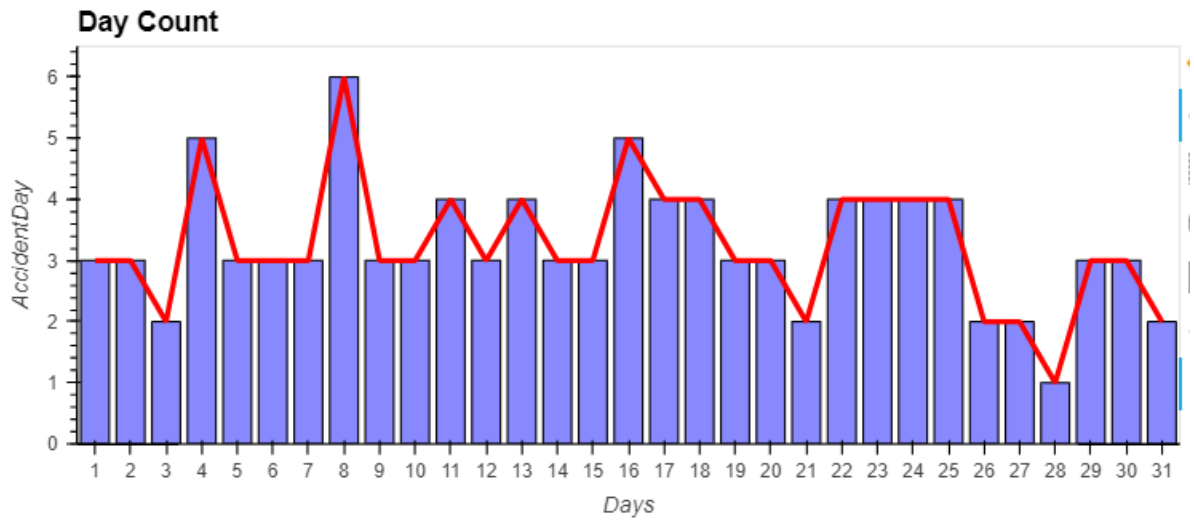

Figure 14: Accidents reported by Day of the Month

Table 13: Accidents reported distribution by Day of the Month

Day of Month	Count	%	Cumulative %
8	25	6.0%	6.0%
4	22	5.3%	11.2%
16	19	4.5%	15.8%
11	18	4.3%	20.1%
22	17	4.1%	24.2%
23	16	3.8%	28.0%
24	16	3.8%	31.8%
13	15	3.6%	35.4%
17	15	3.6%	39.0%
18	15	3.6%	42.6%
25	15	3.6%	46.2%
1	14	3.3%	49.5%
5	14	3.3%	52.9%
15	14	3.3%	56.2%
30	14	3.3%	59.6%
2	13	3.1%	62.7%
6	13	3.1%	65.8%
10	13	3.1%	68.9%
9	12	2.9%	71.8%
14	12	2.9%	74.6%
29	12	2.9%	77.5%
7	11	2.6%	80.1%
12	11	2.6%	82.8%
19	11	2.6%	85.4%
20	11	2.6%	88.0%
3	10	2.4%	90.4%
21	9	2.2%	92.6%

Day of Month	Count	%	Cumulative %
26	9	2.2%	94.7%
27	9	2.2%	96.9%
31	7	1.7%	98.6%
28	6	1.4%	100.0%

- 8th Day of the Month has the highest number of Accidents (25 at 6% of Total), followed by 4th Day (22 at 5.3% of Total), 16th Day (19 at 4.5% of Total), 11th Day (18 at 4.3% of Total), and 22nd Day (17 at 4.1% of Total). The top 5 make up 24.2% of Total Accidents reported.
- 28th Day of the Month has the lowest number of Accidents (6 at 1.4 % of Total)
- While not every month has 31 days, 31 still ranks higher than 28 which occurs in every month, including February. Days 29 and 30 also show a moderately high number of Accidents. This could be due to the urgency or pressure due to month end deadline

3.1.5 Bivariate Analysis

3.1.5.1 Industry Sector and Country

Does the distribution of Accidents reported by Industry Sector differ significantly in different Countries or not?

Table 14: Accidents reported by Industry Sector and by Country

Industry Sector by Country	Country_01	Country_02	Country_03
Metals	46	88	0
Mining	200	37	0
Others	2	4	41

The is no data for Accidents in Mining and Metals sector for Country_03. All of the Accidents reported from Country_03 are classified under Others Industry Sector.

Among Country_01 and Country_02, the distribution of Accidents reported across Mining and Metals is significantly different. Country_02 has more Accidents reported in Metals sector as compared to Mining, while Country_01 has significantly more Accidents reported in mining sector as compared to Metals.

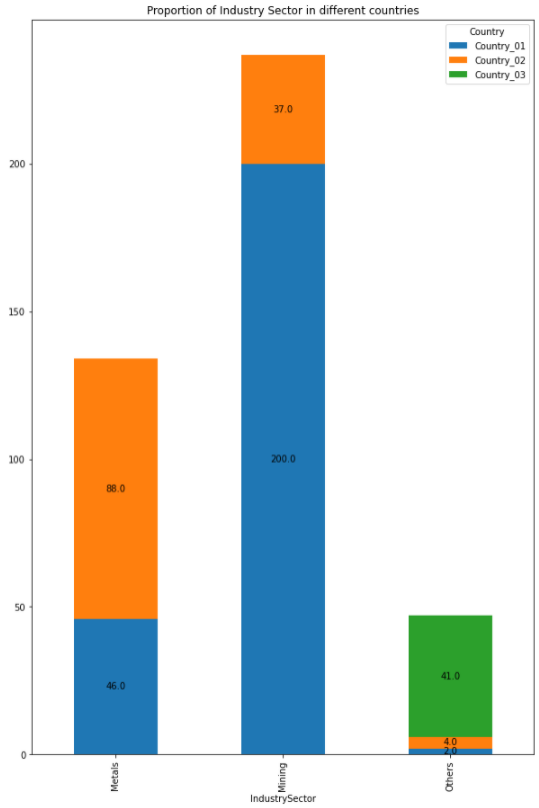

Figure 15: Accidents reported by Industry Sector split by Countries

3.1.5.2 Employee Type by Gender

Is the distribution of Employee Type of the Person involved in Accident reported significantly different for different Genders?

It must be noted that Number of Females were significantly lower that Number of Males in the Accident data. Hence, we look at percentages within each Gender to elicit insights.

Among Males, percentage of Employees and Third Party Personnel getting injured in an Accident is nearly same in the neighborhood of 40% which is also similar to the percentage of Third Party Female Personnel.

When we look at Third Party Personnel who are Remote, Females have a higher percentage of injuries in Accidents than the Males.

Among Employees, percentage of Males is slightly more than Females on injuries in Accidents.

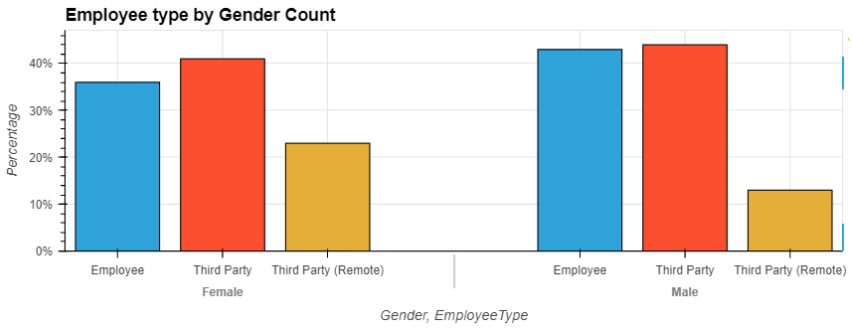

Figure 16: Accident % by Employee Type and Gender

3.1.5.3 Industry Sector by Gender

Does the distribution of Accidents reported by Industry Sector differ significantly with different Genders?

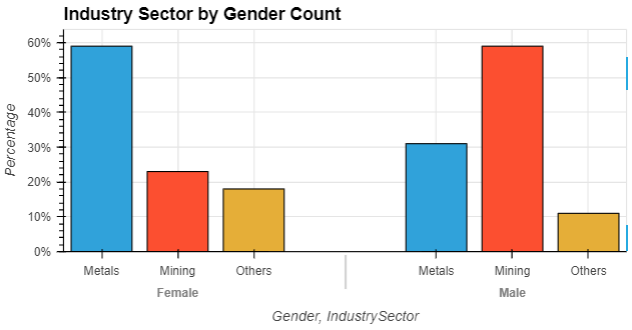

Figure 17: Accident % by Industry Sector and Gender

It must be noted that Number of Females were significantly lower that Number of Males in the Accident data. Hence, we look at percentages within each Gender to elicit insights.

Among Females, the highest proportion of the Accidents were reported in Metals Industry Sector (~60%) as compared to Mining and Others which are both in the neighborhood of ~20%

Among Males, the highest proportion of the Accidents were reported in Mining Industry Sector (~60%) as compared to Metals (~30%) and Others (~10%)

Males in Mining Industry Sector have more than twice the percentage of Accidents reported than Females, while Females in Metals Industry Sector have almost twice the percentage of Accidents reported than Males. In the Others Industry Sector, Females have a higher percentage of Accidents reported than Males.

3.1.5.4 Accident Level by Gender

Does the distribution of Accident Level and Potential Accident Level differ significantly with different genders?

It must be noted that Number of Females were significantly lower that Number of Males in the Accident data. Hence, we look at percentages within each Gender to elicit insights.

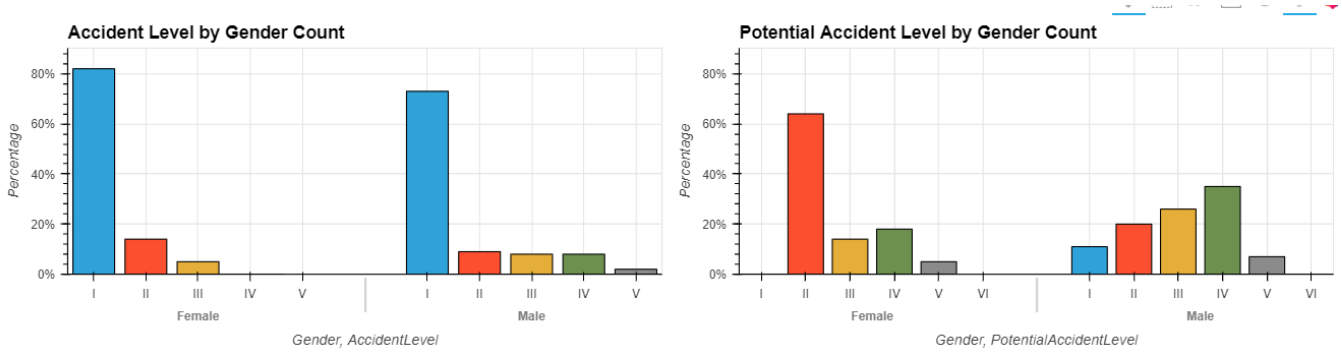

Figure 18: Accident Level by Gender - Reported and Potential

For Females, majority of the Accidents reported were classified with Accident Level I (lowest severity), while majority of the values of Potential Accident Level is II or in some cases higher.

For Males, majority of the Accidents reported were classified with Accident Level I (lowest severity), while the Potential Accident Level is concentrated more towards IV, III, II in that order.

By looking at the distribution of Potential Accident Level, Males are likely at a higher-level risk of Accident and injury than Females by nature of the activities they typically perform.

3.1.5.5 Employee Type by Accident Level

Does the Accident Level and Potential Accident Level differ significantly for different Employee Types?

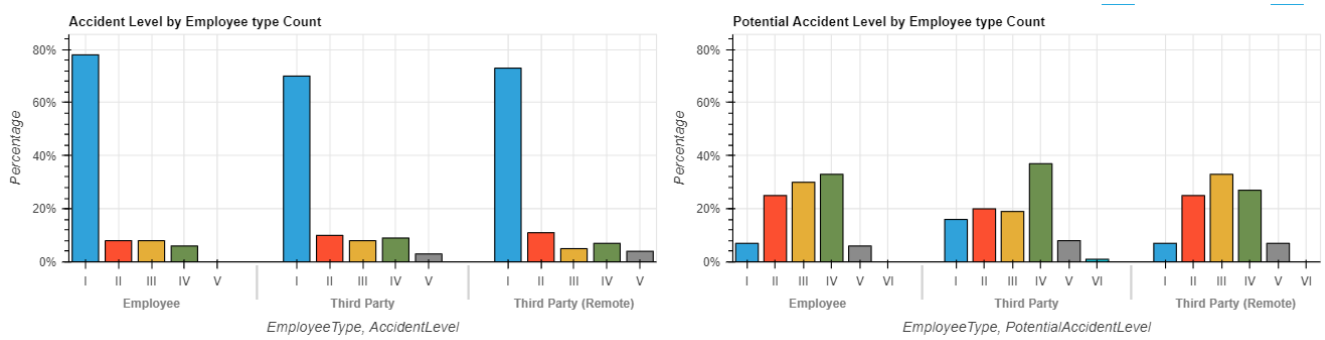

Figure 19: Accident Level by Employee Type - Reported and Potential

Across all the Employee Types, majority of the Accidents are reported at Accident Level I (lowest severity). However, the Potential Accident Level for all Employee Types is more inclined towards Levels IV, III, II.

Third Party personnel and Third Party (Remote) personnel are involved in a higher proportion of Accidents that have a high Potential Accident Level as compared to Employees. That said, even for Employees, a majority of Accidents have a Potential Accident Level in IV, III, II.

Third Party Personnel and Third Party (Remote) Personnel have higher proportion of Accidents reported at Accident Level V (which is one level below the highest severity of VI) as compared to Employees.

Third Party was involved in one Accident that had the highest Potential Accident Level of VI (highest severity). However, no Accident was ever reported at Accident Level VI.

3.1.5.6 Accident Level by Month

Does the distribution of Accident Level and Potential Accident Level differ significantly in different Months?

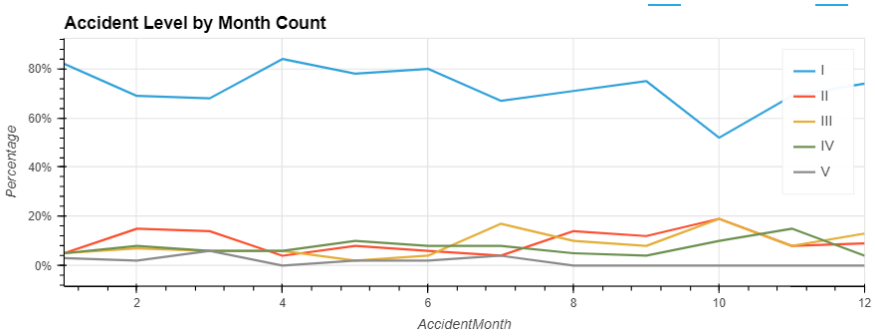

Figure 20: Accident Level by Month – Trendline

Accident Level for mid-level severity Accidents (II, III, IV) is fairly constant through all the months of the year with a slight increase towards the second half of the year. Higher-level severity Accidents (V) have reduced towards the last few months of the year.

Accident Level of I dips sharply in October (10th Month).

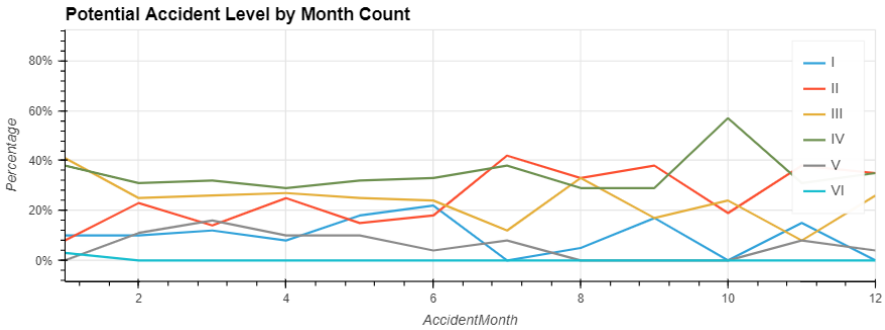

Figure 21: Potential Accident Level by Month – Trendline

Potential Accident Level is increasing towards the second half of the year, particularly for the mid-level severity Accidents (II, III, IV).

Potential Accident Level of III peaks sharply in August (8th Month)

Potential Accident Level of IV peaks sharply in October (10th Month) with proportional dip in I and II.

3.1.5.7 Accident Level by Day of the Week

Does the distribution of Accident Level and Potential Accident Level differ significantly in different Days of the Week?

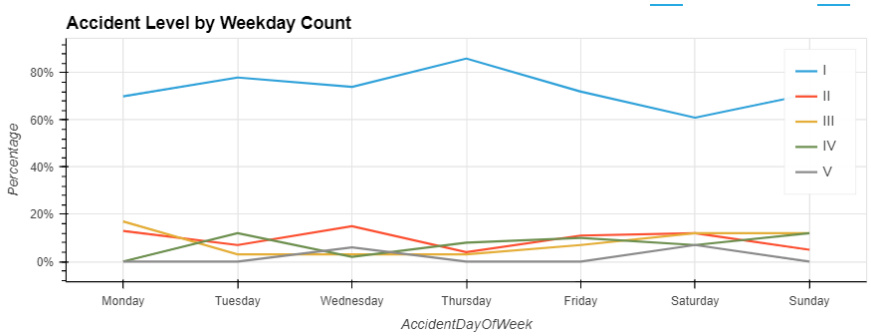

Figure 22: Accident Level by Day of the Week – Trendline

Accident Level is mostly consistent across the Days of the Week with a minor peak on Thursdays for level I (lowest severity) Accidents.

Accident Level of V (one level below the highest severity of VI) shows minor twin-peaks on Wednesdays and Saturdays.

Accident Level of IV shows minor twin-peaks on Tuesdays and Fridays.

Accident Level of I shows minor twin-peaks on Wednesdays and Saturdays.

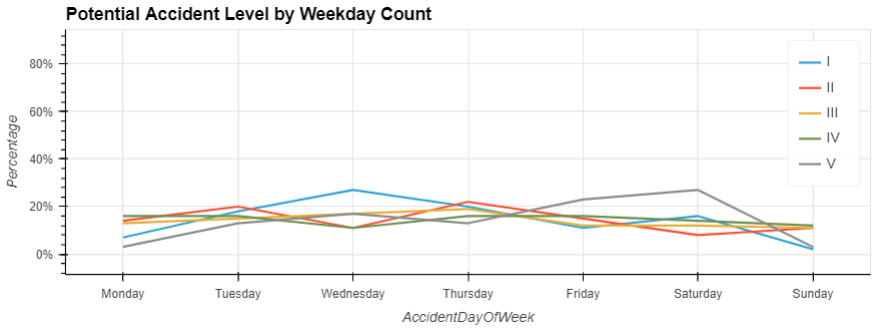

Figure 23: Potential Accident Level by Day of the Week – Trendline

Potential Accident Level is mostly consistent across the Days of the Week with a minor peak on Wednesdays for level I (lowest severity) Accidents.

Accident Level of V (one level below the highest severity of VI) shows minor twin-peaks on Fridays and Saturdays.

3.1.5.8 Accident Level by Season of the Year

Does the distribution of Accident Level and Potential Accident Level differ significantly in different Seasons of the Year?

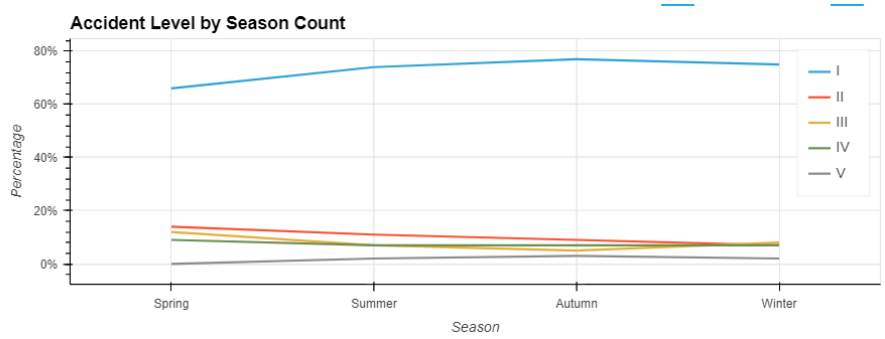

Figure 24: Accident Level by Season of the Year – Trendline

Accident Level is fairly consistent across all Seasons of the Year with a gradual decrease towards the second half of the year, primarily due to lack of data for second half of year 2017.

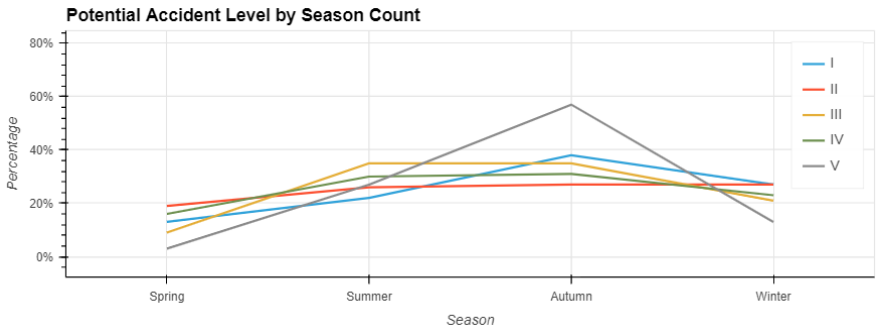

Figure 25: Potential Accident Level by Season of the Year – Trendline

Overall there is a higher level of Potential Accident Level in Autumn, particularly for level V (one level below the highest severity of VI)

3.1.6 Crosstab Analysis

3.1.6.1 Accident Level vs. Potential Accident Level

Potential_Acc_Level	I	II	III	IV	V	VI	Total
AccidentLevel							
I	45	88	89	78	9	0	309
II	0	7	14	16	3	0	40
III	0	0	3	26	2	0	31
IV	0	0	0	21	9	0	30
V	0	0	0	0	7	1	8
Total	45	95	106	141	30	1	418

Figure 26: Crosstab: Accident Level vs. Potential Accident Level

The Potential Accident Level on an Accident reported is usually at the same level of severity or higher than the actual recorded Accident Level.

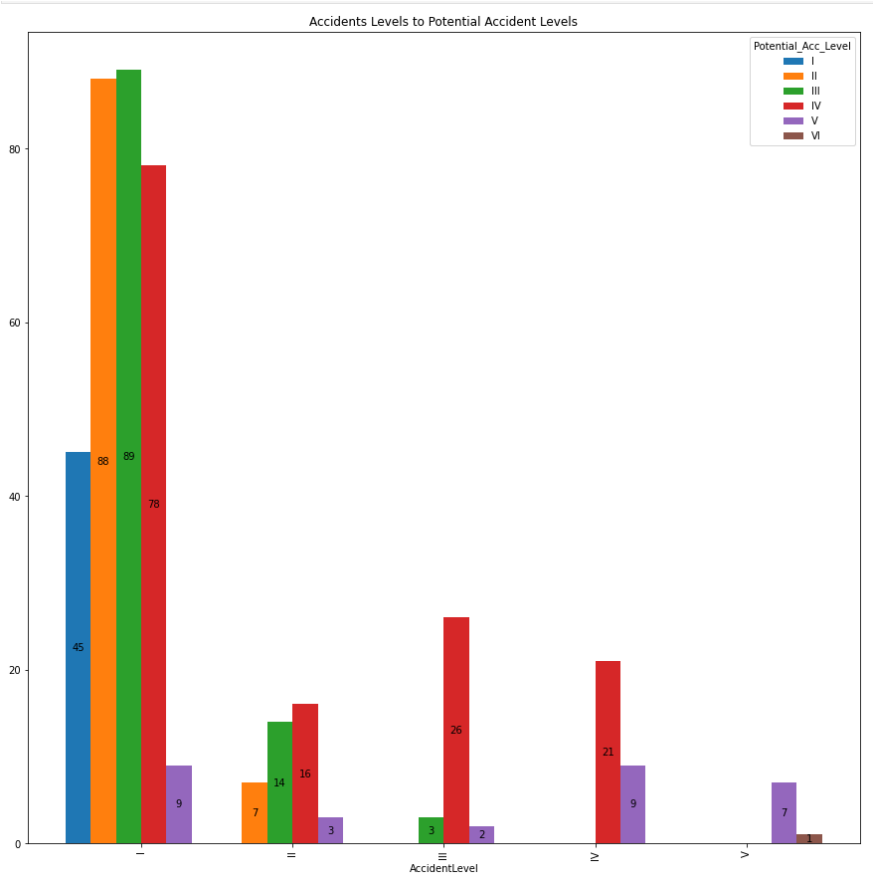

Figure 27: Accident Level vs. Potential Accident Level

Majority of the Accidents reported at Accident Level I (lowest severity) have a higher Potential Accident Level indicating that the Accident could have been more severe and the Industry overall has higher risk activities.

A significant number of Accidents reported at Accident Level III have a Potential Accident Level of IV

Level IV is where there is a most alignment of both the reported Accident Level and the Potential Accident Level.

3.1.6.2 Countries vs. Accident Levels

AccidentLevel	I					II				III			IV		V	Total
Potential_Acc_Level	I	II	III	IV	V	II	III	IV	V	III	IV	V	IV	V	V	VI	
Country																	
Country_01	10	50	55	58	4	1	7	9	2	2	17	2	17	6	7	1	248
Country_02	6	36	34	17	5	4	7	7	1	0	7	0	2	3	0	0	129
Country_03	29	2	0	3	0	2	0	0	0	1	2	0	2	0	0	0	41
Total	45	88	89	78	9	7	14	16	3	3	26	2	21	9	7	1	418

Figure 28: Crosstab: Country vs. Accident Levels - Reported and Potential

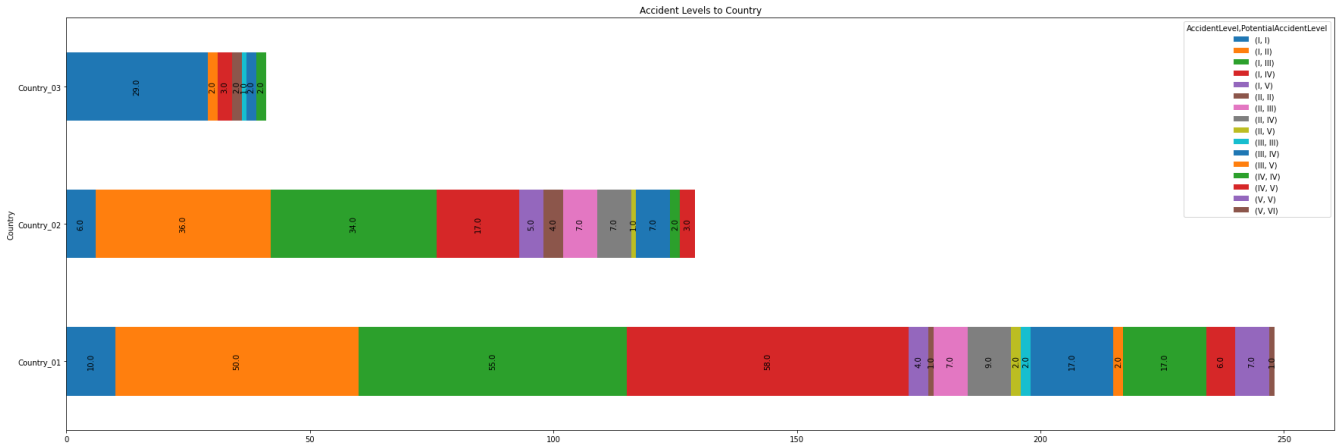

Figure 29: Country vs. Accident Levels - Reported and Potential

Country_03 has most of the Accidents reported where both Accident Level and Potential Accident Level are at the lowest severity of I. the proportion of other Accident levels is significantly lower.

Country_02 and Country_03 have a vast majority of instances where the reported in Accident Level is at lowest severity of I, while the Potential Accident Level is spread across II, III, IV and on occasions spilling into V as well.

Country_01 will have to particularly investigate more on whether the reported Accident Levels were indeed correctly classified at Level I or not, since a vast majority of the Accidents fall under Potential Accident Level of III, IV, V. Country_02 also has a similar observation on this aspect even though it not as glaring as Country_01.

When Countries perform further analysis, they may also want to drill down further into the Locals in their Country to either isolate the anomalies or understand the trends.

A summarized view of Accident Level and Potential Accident Level by Local is also shown below to serve as a reference.

AccidentLevel	I					II				III		IV		V		Total
Potential_Acc_Level	I	II	III	IV	V	II	III	IV	V	III	IV	V	IV	V	V	VI	
Local																	
Local_01	2	7	12	23	1	0	1	0	0	0	4	1	3	1	1	0	56
Local_02	1	4	4	2	3	2	2	2	0	0	2	0	1	0	0	0	23
Local_03	2	12	21	27	3	0	0	6	2	0	4	1	4	4	2	1	89
Local_04	2	10	13	5	0	1	5	3	0	1	6	0	5	1	3	0	55
Local_05	1	15	23	12	0	0	2	4	0	0	2	0	0	0	0	0	59
Local_06	4	20	9	3	0	0	1	0	0	1	2	0	5	0	1	0	46
Local_07	1	4	1	1	2	0	1	0	1	0	1	0	0	2	0	0	14
Local_08	2	10	6	1	0	2	2	1	0	0	1	0	1	1	0	0	27
Local_09	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
Local_10	29	2	0	3	0	2	0	0	0	1	2	0	2	0	0	0	41
Local_11	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
Local_12	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	4
Total	45	88	89	78	9	7	14	16	3	3	26	2	21	9	7	1	418

Figure 30: Crosstab: Local vs Accident Level - Reported and Actual

3.1.6.3 Industry Sector vs. Accident Level

AccidentLevel	I					II					III		IV			V	Total
Potential_Acc_Level	I	II	III	IV	V	II	III	IV	V	III	IV	V	IV	V	V	VI	
IndustrySector																	
Metals	7	46	38	16	0	2	5	5	0	1	6	0	6	1	1	0	134
Mining	8	37	51	58	9	3	9	11	3	1	17	2	13	8	6	1	237
Others	30	5	0	4	0	2	0	0	0	1	3	0	2	0	0	0	47
Total	45	88	89	78	9	7	14	16	3	3	26	2	21	9	7	1	418

Figure 31: Crosstab: Industry Sector vs. Accident Level - Reported and Potential

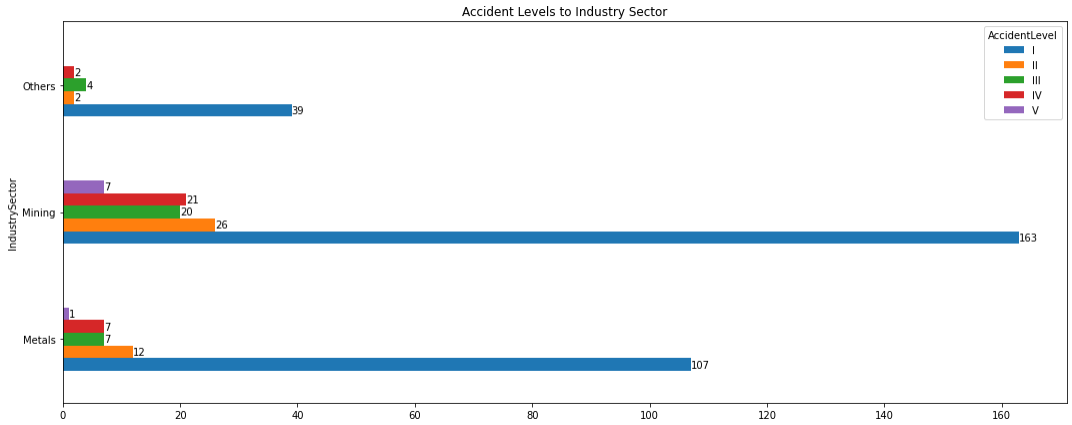

Figure 32: Industry Sector vs. Accident Level - Reported and Potential

As consistent with earlier analysis, across the Industry Sectors, a majority of the Accidents are reported with Accident Level of I (lowest severity)

3.1.6.4 Gender vs. Accident Level

AccidentLevel	I					II					III		IV			V	Total
Potential_Acc_Level	I	II	III	IV	V	II	III	IV	V	III	IV	V	IV	V	V	VI	
Gender																	
Female	0	13	2	2	1	1	1	1	0	0	1	0	0	0	0	0	22
Male	45	75	87	76	8	6	13	15	3	3	25	2	21	9	7	1	396
Total	45	88	89	78	9	7	14	16	3	3	26	2	21	9	7	1	418

Figure 33: Crosstab: Gender vs. Accident Level - Reported and Potential

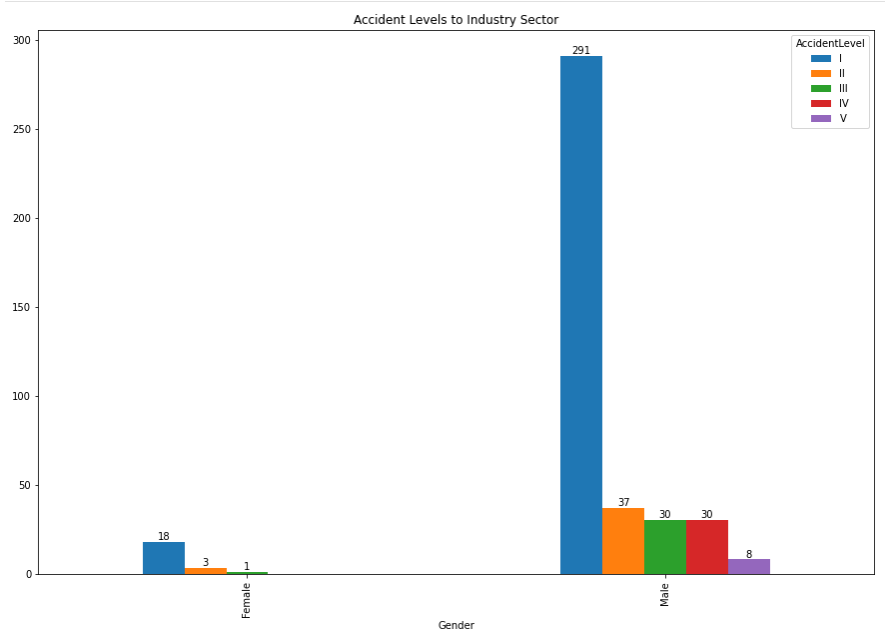

Figure 34: Gender vs. Accident Level

As consistent with earlier analysis, across the Genders, a majority of the Accidents are reported with Accident Level of I (lowest severity)

3.1.6.5 Employee Type vs. Accident Level

AccidentLevel	I					II				III		IV		V	Total
Potential_Acc_Level	I	II	III	IV	V	II	III	IV	V	III	IV	V	IV	V	VI	
EmployeeType																
Employee	12	43	44	35	5	1	6	6	2	3	11	0	6	4	0	178
Third Party	29	34	29	35	3	3	6	9	1	0	12	2	12	4	5	1	185
Third Party (Remote)	4	11	16	8	1	3	2	1	0	0	3	0	3	1	2	0	55
Total	45	88	89	78	9	7	14	16	3	3	26	2	21	9	7	1	418

Figure 35: Crosstab: Employee Type vs. Accident Level - Reported and Potential

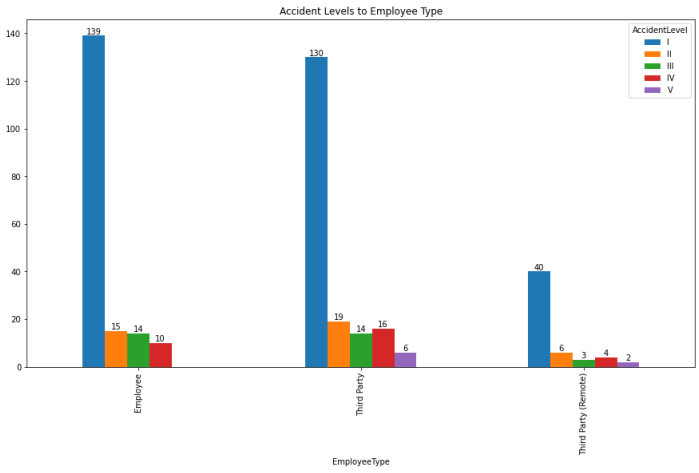

Figure 36: Employee Type vs. Accident Level - Reported and Actual

As consistent with earlier analysis, across the Employee Types, a majority of the Accidents are reported with Accident Level of I (lowest severity)

3.1.6.6 Seasons vs. Accident Level

AccidentLevel	I					II					III		IV			V	Total
Potential_Acc_Level	I	II	III	IV	V	II	III	IV	V	III	IV	V	IV	V	V	VI	
Season																	
Autumn	17	24	34	26	7	2	3	5	2	0	5	2	8	2	4	0	141
Spring	6	16	6	10	0	2	3	3	0	1	6	0	4	1	0	0	58
Summer	10	24	30	25	2	1	5	6	1	2	7	0	4	4	1	1	123
Winter	12	24	19	17	0	2	3	2	0	0	8	0	5	2	2	0	96
Total	45	88	89	78	9	7	14	16	3	3	26	2	21	9	7	1	418

Figure 37: Seasons vs. Accident Level - Reported and Potential

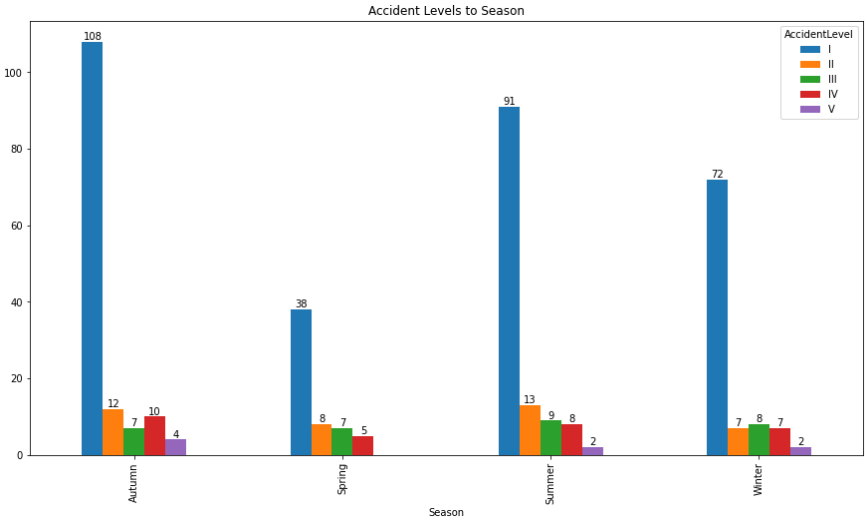

Figure 38: Seasons vs. Accident Level - Reported and Potential

As consistent with earlier analysis, across the Seasons, a majority of the Accidents are reported with Accident Level of I (lowest severity)

3.1.6.7 Holidays vs. Accident Level

AccidentLevel	I					II					III		IV			V	Total
Potential_Acc_Level	I	II	III	IV	V	II	III	IV	V	III	IV	V	IV	V	V	VI	
IsHoliday																	
0	42	88	89	73	8	7	14	15	3	3	26	2	21	9	7	1	408
1	3	0	0	5	1	0	0	1	0	0	0	0	0	0	0	0	10
Total	45	88	89	78	9	7	14	16	3	3	26	2	21	9	7	1	418

Figure 39: Crosstab: Holiday vs. Accident Level - Reported and Potential

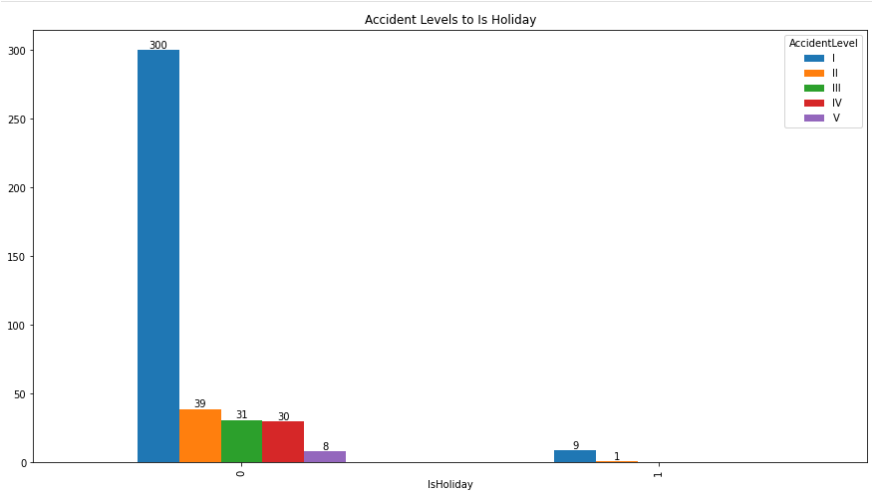

Figure 40: Holidays vs. Accident Level

On a Holiday, there haven't been any significant number of Accidents reported to have any meaningful insights.

3.2 Solution Architecture Evaluation

3.2.1 NLP Pre-Processing of Data

The Description column in the dataset contains the natural language text that is pre-processed to train the Chatbot.

The following sequence of pre-processing transformations were performed:

Table 15: NLP Pre-Processing of Description Data

Pre-Processing Step	Purpose	Method Used	Output
Remove Stop Words A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query	The Chatbot would not find anything useful from the stop words as they are noise that is best eliminated.	from nltk.corpus import stopwords stopwords.words('english') method	The stop words found in the Description are ignored and the other words are picked up for creating word tokens
Tokenize Words Tokenizing involves splitting sentences and words from the body of the text	Extract entities of interest from the text after stop words are removed	from nltk.tokenize import word_tokenize word_tokenize method	The word tokens are extracted from the Description
Lemmatize Words Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Examples: • **rocks : rock** • **corpora : corpus** • **better : good**	Lemmatize words to replace words with similar meanings with one word that retains the context. Lemmatization helps train the Chatbot to handle multiple variations of a word a	from nltk.stem import WordNetLemmatizer lemmatize method	The word tokens are lemmatized

Pre-Processing Step	Purpose	Method Used	Output
	user might use in the text input		
Ignore Punctuations	Punctuations serve no purpose for extracting the context and intent for the Chatbot	List of punctuations from string.punctuation	Punctuations are excluded from the text
Convert to Lower Case	Since python is case-sensitive, it is best to convert all words to lower case for consistency and simplifying training		
Ignore Numeric Text	Ignore numeric text and pick up only alphabetic text		

After NLP Pre-Processing, the lengths of the Description text were analyzed

- Minimum line length was 61 characters
- Maximum line length was 664 characters

Line with maximum length is: level gallery holding activity bolter equipment operator performs drilling first hole support right gable when drill end drill rod break leaving thread inside drilling machine shank operator assistant decide make two empty percussion attempt free thread shank without success third attempt assistant enters corrugated iron central hole rest bar embedded shank generate pressure moment operator activates percussion generates movement shank hit palm victim left hand generating described injury the worker wearing safety glove time accident the end corrugated iron contact left hand shaped like cane the worker time accident positioned roof supported mesh split set

Example of how NLP Pre-processing has transformed the Description text:

> **Description text before NLP Pre-Processing:** "While removing the drill rod of the Jumbo 08 for maintenance, the supervisor proceeds to loosen the support of the intermediate centralizer to facilitate the removal, seeing this the mechanic supports one end on the drill of the equipment to pull with both hands the bar and accelerate the removal from this, at this moment the bar slides from its point of support and tightens the fingers of the mechanic between the drilling bar and the beam of the jumbo"
>
> **Description text after NLP Pre-Processing:** "while removing drill rod jumbo maintenance supervisor proceeds loosen support intermediate centralizer facilitate removal seeing mechanic support one end drill equipment pull hand bar accelerate removal moment bar slide point support tightens finger mechanic drilling bar beam jumbo"

The number of words range from 10 to 97.

Line with minimum number of words: check list area survey operator slipped foliage leucenas fell

Line with maximum number of words: level gallery holding activity bolter equipment operator performs drilling first hole support right gable when drill end drill rod break leaving thread inside drilling machine shank operator assistant decide make two empty percussion attempt free thread shank without success third attempt assistant enters corrugated iron central hole rest bar embedded shank generate pressure moment operator activates percussion generates movement shank hit palm victim left hand generating described injury the worker wearing safety glove time accident the end corrugated iron contact left hand shaped like cane the worker time accident positioned roof supported mesh split set

3.2.1.1 Length of Description (Characters)

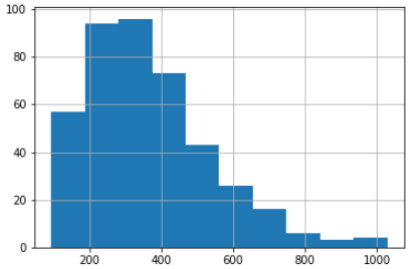

Figure 41: Histogram of Length of Description before NLP Pre-Processing

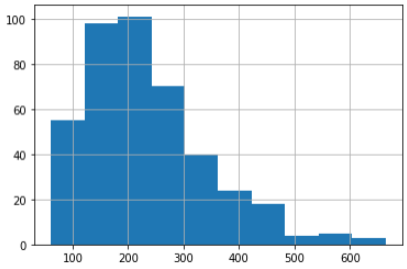

Figure 42: Histogram of Length of Description after NLP Pre-Processing

3.2.1.2 Word Count in Description

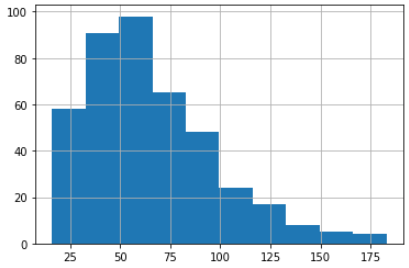

Figure 43: Histogram of Number of Words in Description before NLP Pre-Processing

A comparison of the before and after histograms reveals that nearly 50% of the words in the Descriptions, on an average, have been eliminated through the various steps of NLP Pre-Processing to extract the entities that are of significance for NLP

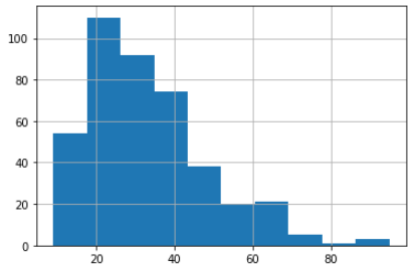

Figure 44: Histogram of Number of Words in Description after NLP Pre-Processing

3.2.1.3 Average Length of Words

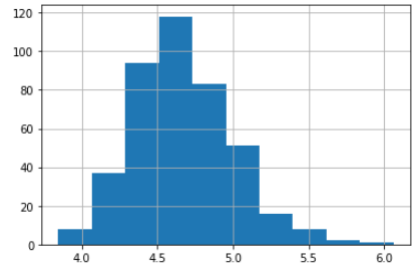

Figure 45: Average Word Length in Description before NLP Pre-Processing

Average world length has increased due to elimination of shorter words like stop words.

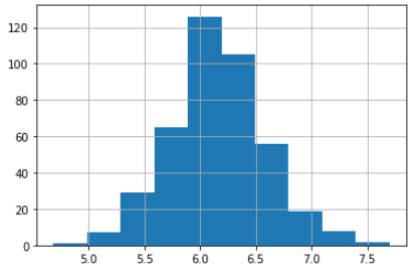

Figure 46: Average Word Length in Description after NLP Pre-Processing

3.2.1.4 N-Grams

In natural language processing n-gram is a contiguous sequence of n items generated from a given sample of text where the items can be characters or words and n can be any numbers like 1,2,3, etc. N-Grams are useful to create features from text corpus for machine learning algorithms like SVM, Naive Bayes, etc. N-Grams are useful for creating capabilities like autocorrect, autocompletion of sentences, text summarization, speech recognition, etc.

The following figures summarize the N-Grams for the Accident Description text after NLP Pre-Processing:

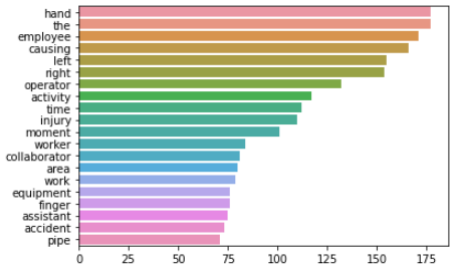

Figure 47: Frequency of Word Sequences N-Gram (N=1)

In the Unigram (N-Gram where N = 1), the word "hand" tops the list of most frequently occurring words in the dataset for the Description column, along with "the", followed by "employee", "causing", "left", "right", "operator".

An inspection of the Bigram (N=2) and Trigram (N=3) would reveal more insights on what word sequences are frequently occurring in the Description text.

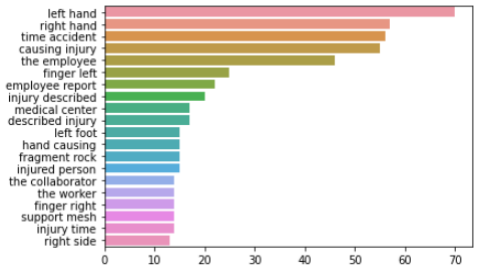

Figure 48: Frequency of Word Sequences N-Gram (N=2)

The Bigram (N-Gram with N=2) is more insightful and has brought together pairs of words in a sequence that are most commonly occurring in the Description text.

The word "hand" still tops the list with better qualification on which hand, with "left hand" most frequently mentioned, followed by "right hand". A useful insight that can probably be gathered for better root cause analysis is asking the question on "What is the dominant hand of the person involved in the accident? Are they left-handed, or right-handed, or ambidextrous?". It may also be worth checking if the person was already nursing an injury or had any limitation or disability. In addition to hands, "finger left", "finger right", and "left foot" are other human body parts that are mentioned more often in the Description of the Accident.

References to the injured person as "the employee", "the collaborator", "the worker" show up in the list of frequently occurring bigrams.

The word sequence "fragment rock" stands out as the sole instance of what object most frequently is the cause of the Accident and Injury, most likely from the Mining sector.

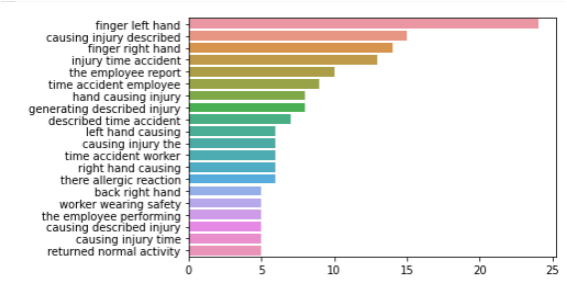

Figure 49: Frequency of Word Sequences N-Gram (N=3)

The Trigram (N-Gram with N=3) further accentuates the most frequently occurring word sequences which give a lot more context, with "finger left hand" featuring at the top of the list by a significant margin from the rest of the word sequences and "finger right hand" is at third position.

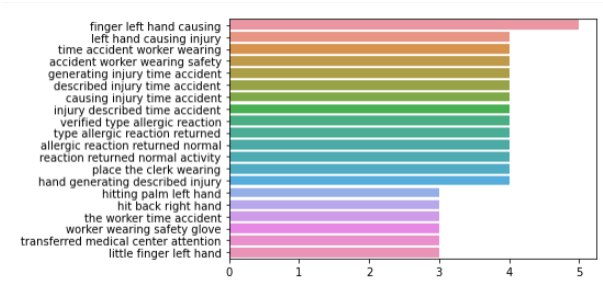

Figure 50: Frequency of Word Sequences N-Gram (N=4)

At N=4, the frequencies of word sequences are no longer very distinctive or varied and aren't any more insightful.

3.2.1.5 Word Cloud

Figure 51: Word Cloud of Description before NLP-Preprocessing

Figure 52: Word Cloud of Description after NLP-Preprocessing

3.2.1.6 Five-Point Summary of Words

Table 16: NLP-Preprocessing - Five-Point Summary of Words in Description

Count of Descriptions	418 Lines
Mean No. of Words	33
Standard Deviation of No. of Words	15
Minimum No. of Words	9
25th Percentile	21
50th Percentile	30
75th Percentile	41
99th Percentile	76
Maximum No. of Words	95

Additional insights on Length of Accident Description and corresponding Accident Level classification

- 74% of data where accident description > 100 is captured in low accident level.
- 34% of data where accident description > 100 is captured in high - medium potential accident level.
- 25% of data where accident description > 100 is captured in medium potential accident level.
- 23% of data where accident description > 100 is captured in low potential accident level.

3.2.2 Featurization, Model Selection & Tuning Strategy

The current status as on 12 Dec 2021 as part of the interim report. This section will be revised in the final report with the decision taken and the reason for the decision.

3.2.2.1 Featurization

Choices for Target Variables are being evaluated and will be decided based on the Model performance:

- Only “Accident Level” predicted by a Model trained on Input Dataset

- Both “Accident Level” and “Potential Accident level” predicted by two respective Models (one for each Target Variable) trained on Input Dataset
- Both “Accident Level” and “Potential Accident Level” combined into a single composite Target Variable “Accident Level – Potential Accident Level” (with values such as “I – II”, “IV - V” etc.) predicted by a Model trained on Input Dataset

Choices for the Input Variables would be evaluated and decided based on Model performance and Data Analysis. Evidently, the Accident “Description” text is an input that influences the Accident Level classification. However, other inputs such as Country, Local, Industry Type, Employee Type etc. are being evaluated.

Gender data was too skewed to be able to influence the outcome.

Critical Risk data was too skewed to be able to influence the outcome as most of them were grouped under “Others”.

While the Date of the Accident including the derived data points on Year, Month, Day of the Month, Day of the Week, Season, Holiday etc. were useful to analyze the trend of Accidents reported, there weren’t any statistical significance to warrant inclusion as inputs to the model.

Thus, the decision to be taken will be on what will be the input variable(s):

- Only the Accident “Description” text
- In addition to Accident “Description” text, other relevant Categorial input variables

The choice if inputs to the Model will also impact the design of the Chatbot to ensure the User Input is gathered for fetching the responses from the model on Accident Level classification, which is the primary purpose of the Chatbot.

Chatbot will request the User to provide the “Description” of the Accident. In addition, if other categorial variables are required as input, Chatbot will engage with the User to ask them to select the appropriate value from a list of values.

3.2.2.2 Model Selection & Tuning

A host of Classification Machine Learning Models are being evaluated to identify the best-fit model for further hyperparameter tuning and final model selection, including:

- LogisticRegression
- RidgeClassifier
- KNeighborsClassifier
- SVC
- RandomForestClassifier
- BaggingClassifier
- ExtraTreesClassifier
- AdaBoostClassifier
- GradientBoostingClassifier
- CatBoostClassifier

- LGBMClassifier
- XGBClassifier
- DecisionTreeClassifier

A list of Deep Learning Models are being evaluated to identify the best-fit model in comparison to the Machine Learning Models, including:

- CNN (Convolutional Neural Networks)
- RNN (Recurrent Neural Networks)
- LSTM (Long Short-Term Memory) Recurrent Neural Networks

3.2.2.3 Initial Results Summary

The initial evaluation of Machine Learning Models is summarized below:

	Method	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score	Multi-Class Logloss
1	LogisticRegression	0.76	0.67	0.54	0.67	0.59	0.95
2	RidgeClassifier	0.76	0.70	0.55	0.70	0.62	1.00
3	KNeighborsClassifier	0.76	0.73	0.64	0.73	0.64	6.50
4	SVC	0.74	0.74	0.54	0.74	0.63	0.92
5	DecisionTreeClassifier	0.99	0.55	0.57	0.55	0.56	15.62
6	RandomForestClassifier	0.96	0.74	0.59	0.74	0.65	3.47
7	BaggingClassifier	0.97	0.71	0.55	0.71	0.62	2.42
8	ExtraTreesClassifier	0.99	0.74	0.55	0.74	0.63	1.60
9	AdaBoostClassifier	0.74	0.75	0.61	0.75	0.67	1.35
10	GradientBoostingClassifier	0.95	0.74	0.65	0.74	0.69	0.92
11	CatBoostClassifier	0.99	0.75	0.60	0.75	0.66	0.95
12	LGBMClassifier	0.99	0.73	0.64	0.73	0.68	1.16
13	XGBClassifier	0.83	0.73	0.59	0.73	0.65	0.94

Figure 53: Initial Results from Classification ML Models

Among the ML Models AdaBoostClassifier with an F1-Score of 67% is the best-fit considering how the other models with higher F1-Score are over-fitting the training data

3.2.3 Chatbot Architecture Evaluation

For the Project, we are building a Python-based Chatbot that is the primary solution output. In addition, we are also exploring Rasa platform Chatbot as a stretch goal.

Presently the focus in on the Python-based Chatbot, while exploration is in progress on Rasa-based Chatbot.

3.2.3.1 Python-based Chatbot

The Chatbot developed for the Project implementation is a combination of Rule-based and AI-based Chatbot. The generic interactions between the Chatbot and User is Rule-based responses, while the core purpose of the Chatbot is NLP based understanding of the Description of the Accident inputted by the User and providing the appropriate classification on the Accident Level.

The components of the Python-based Chatbot are:

- **Intents.json:** Intents file houses the Chatbot structure detailing the various tags and their respective responses
- **Rule-based Functions:** Functions related to greeting, goodbye, thanks, no answer
- **Trained LSTM Model:** Fully trained LSTM model which classifies the Accident Level based on the Description provided by the User
- **Deployment on Flask:** Model is deployed on to flask server which enables http end-point which the UI can consume
- **Tkinter User Interface (UI):** User Interface where the user interacts with the Chatbot, inputs the Description and receives the appropriate response

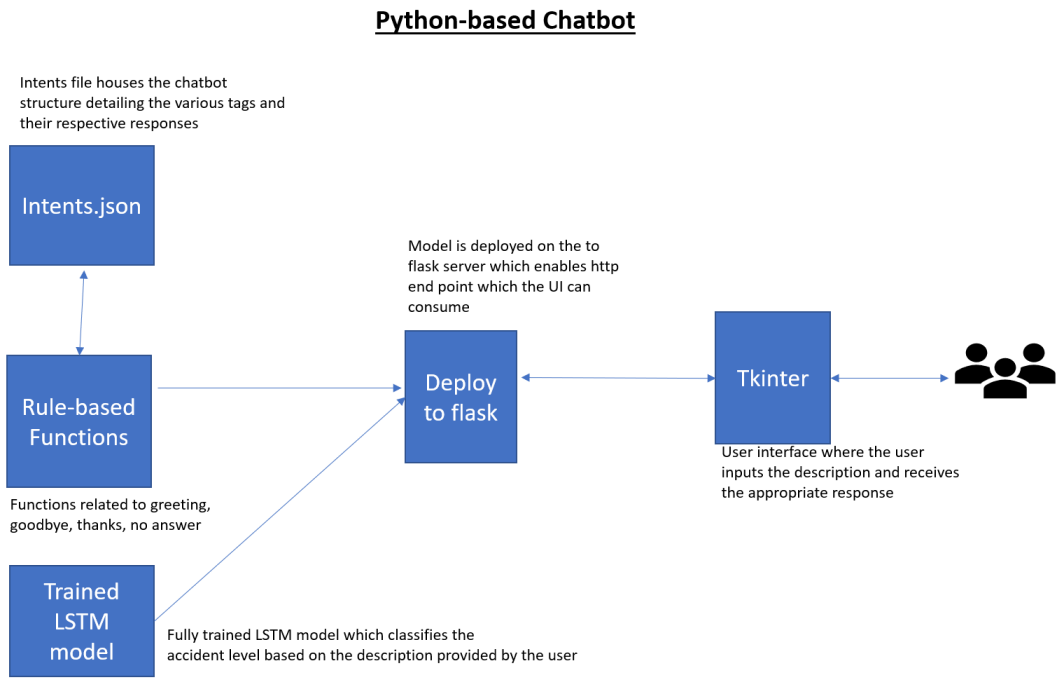

Figure 54: Python-based Chatbot Architecture

This concludes the draft for the Interim Report

Next Steps broadly are on:

Enhancing some of the EDA outputs and creating additional views to take decisions
Finalizing the Input Variables (Predictors) and Target Variables as outlined earlier
Finalizing the Best-Fit Model for the Target Variables and Hyperparameter Tuning
Finalizing the User Interface and User Interaction Script for the Chatbot
Integrating the Model with the Chatbot
In addition to Python-based Chatbot, explore Rasa platform based Chatbot

Chatbot Interface using NLP

4 MODEL TRAINING AND TESTING

List of Models, Parameters, Metrics, and Final Model Pickled

Table 17: Model Selection - Comparison of Metrics

Model					

5 CHATBOT TRAINING AND TESTING

Custom developed Python-based Chatbot

Rasa platform Chatbot

6 PROJECT REPORT

Summary of Results

Screenshot of Chatbot Interaction

Link to Video of Chatbot Interaction

7 PROJECT RETROSPECT

Lessons Learnt

Future Enhancements or Next Steps

Etc.

8 REFERENCES

Chatbot: What is a Chatbot? Why are Chatbots Important? https://www.expert.ai/blog/chatbot/

ELIZA: a Historical Natural Language Processing computer program https://en.wikipedia.org/wiki/ELIZA

The Ultimate Guide to Chatbots https://www.drift.com/learn/chatbot/

Making Sense of the Chatbot and Conversational AI Platform Market https://www.gartner.com/en/documents/3993709/making-sense-of-the-chatbot-and-conversational-ai-platfo

10 Must-Have Chatbot features by Engati https://www.engati.com/blog/10-killer-chatbot-features-business?utm_content=10-killer-chatbot-features-business