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 Attre, 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 Chatbots

These Chatbots follow 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 Chatbots are primarily used by Sales & Sales Development teams. They 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.

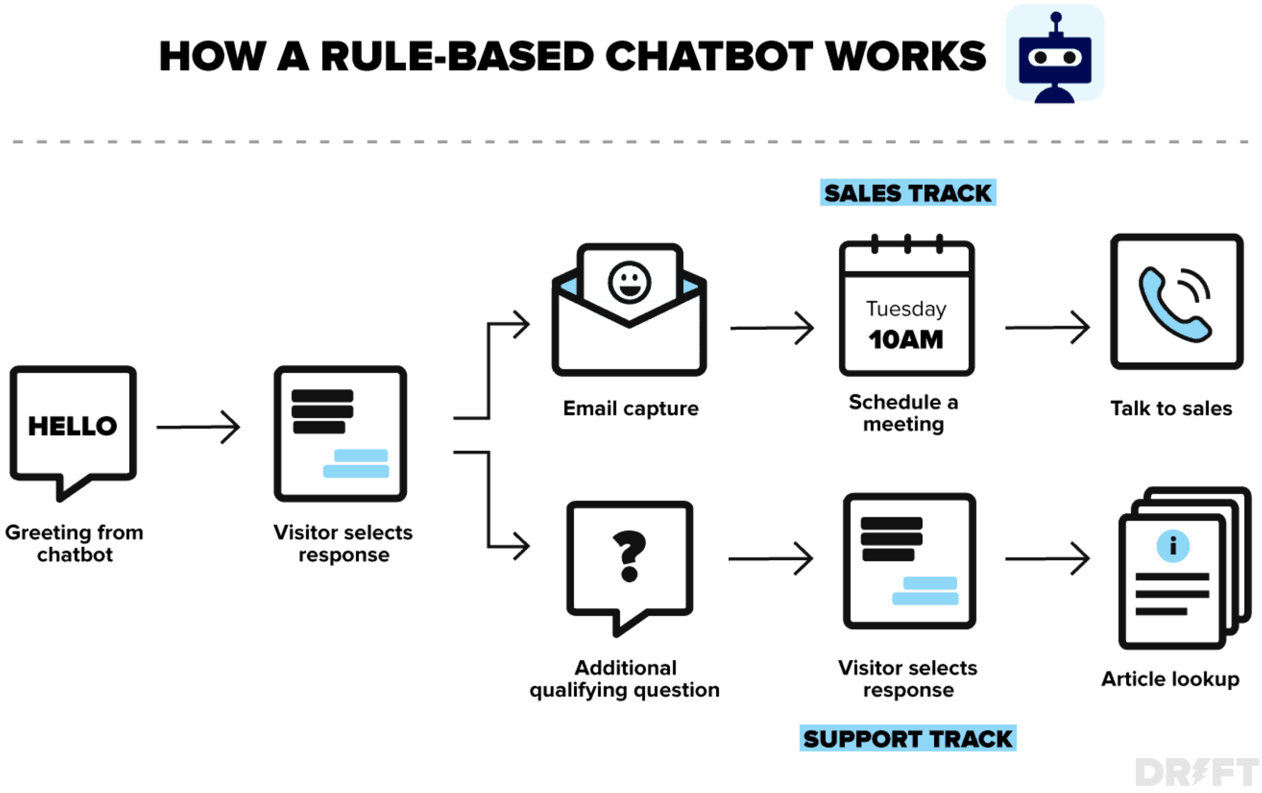


Figure 1: How a Rule-Based Chatbot Works

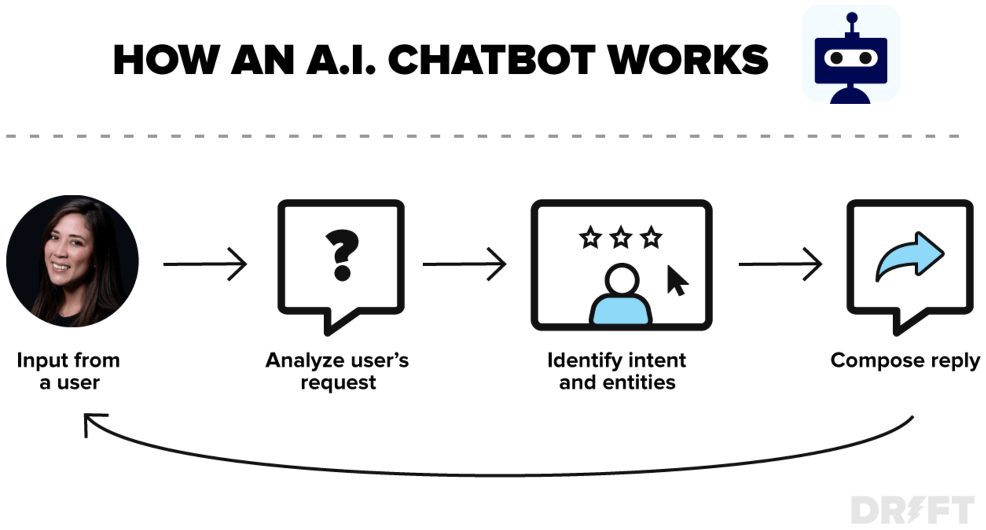


Figure 2: How an AI Chatbot Works

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.

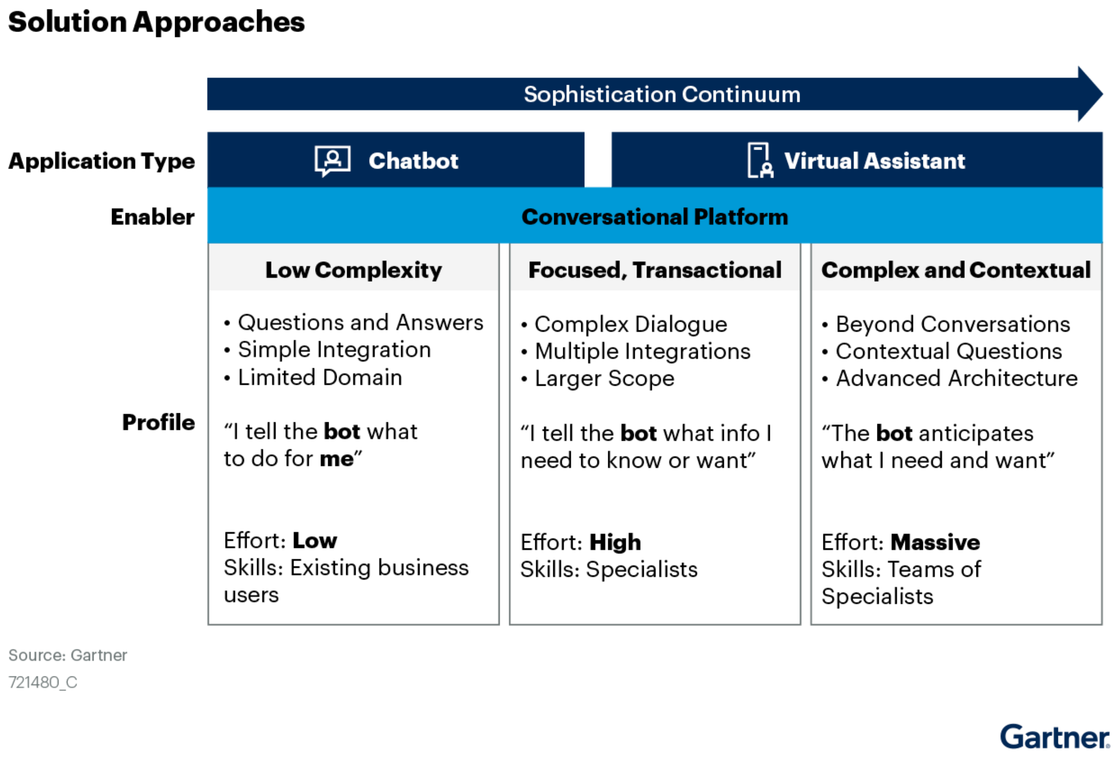
To be successful, a Chatbot solution should be able to effectively perform both of these tasks. Human support plays a key role here: regardless of the kind of approach and the platform, human intervention is crucial in configuring, training and optimizing the Chatbot system.

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.



Solution Approaches

Sophistication Continuum

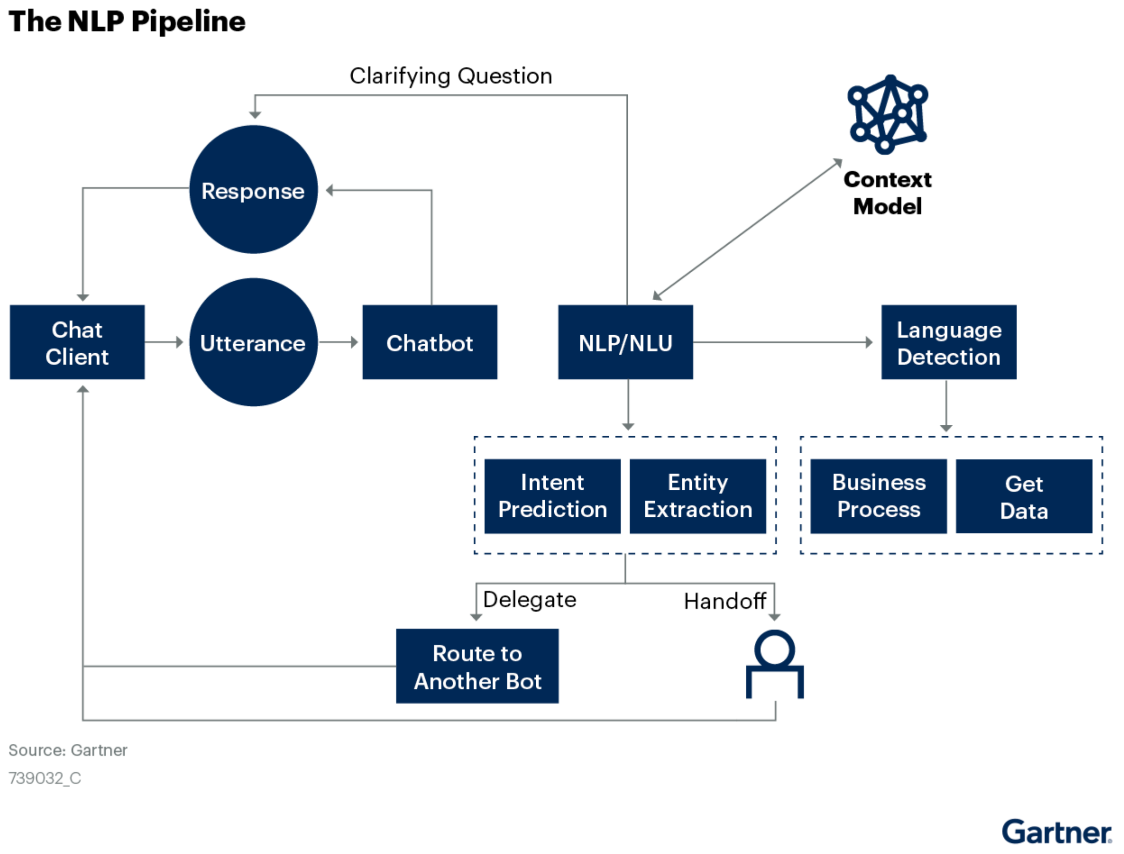
Application Type	Chatbot	Virtual Assistant	
Enabler	Conversational Platform		
	Low Complexity	**Focused, Transactional**	**Complex and Contextual**
Profile	• Questions and Answers • Simple Integration • Limited Domain "I tell the **bot** what to do for **me**" Effort: **Low** Skills: Existing business users	• Complex Dialogue • Multiple Integrations • Larger Scope "I tell the **bot** what info I need to know or want" Effort: **High** Skills: Specialists	• Beyond Conversations • Contextual Questions • Advanced Architecture "The **bot** anticipates what I need and want" Effort: **Massive** Skills: Teams of Specialists

Source: Gartner
721480_C

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

Chatbot Interface using NLP

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.



The NLP Pipeline

Source: Gartner
739032_C

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

2.1.4 Acknowledgments

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 Attre

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 1: 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
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)

Column Name	Column Description	Column Data Type	Initial Observations
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 2: 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 and look for patterns in the data through hypothesis testing, following computed columns were added to the 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): AccidentLevel, PotentialAccidentLevel
- Input Variable(s): AccidentDate, Country, Local, IndustrySector, Gender, EmployeType, CriticalRisk, Description

3.1.4 Univariate Analysis

3.1.4.1 Country

Table 3: Accidents reported by Country

Country	Count	%
Country_01	251	59%
Country_02	130	31%
Country_03	44	10%

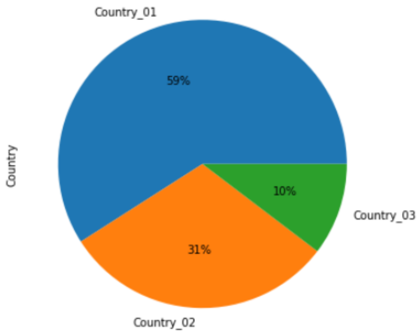


Figure 5: Accidents reported by Country

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

3.1.4.2 Local

Table 4: Accidents reported by Local

Local	Count	%	Cumulative %
Local_03	90	21.2%	21.2%
Local_05	59	13.9%	35.1%
Local_01	57	13.4%	48.5%
Local_04	56	13.2%	61.6%
Local_06	46	10.8%	72.5%
Local_10	44	10.4%	82.8%
Local_08	27	6.4%	89.2%
Local_02	24	5.6%	94.8%
Local_07	14	3.3%	98.1%
Local_12	4	0.9%	99.1%
Local_11	2	0.5%	99.5%
Local_09	2	0.5%	100.0%

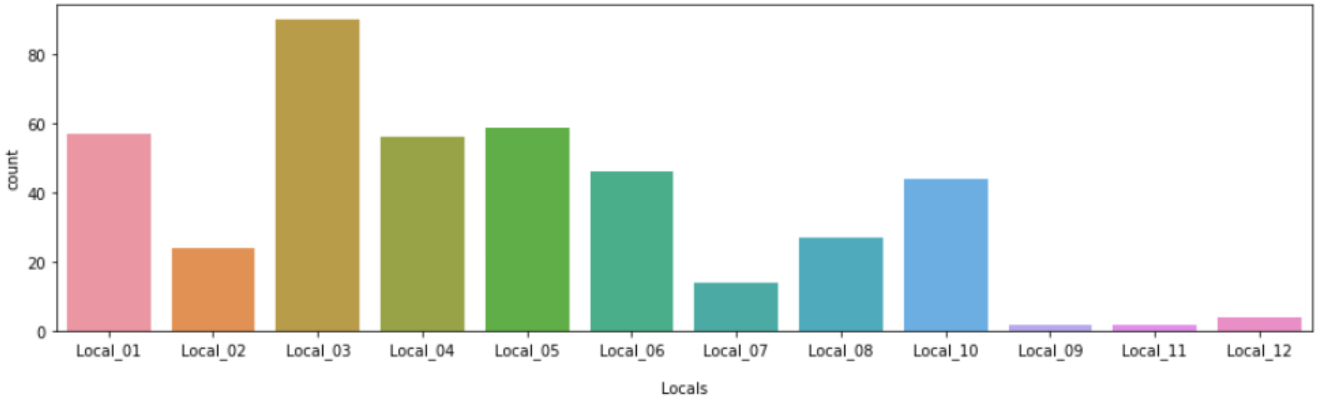


Figure 6: Accidents reported by Local - Counts

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.

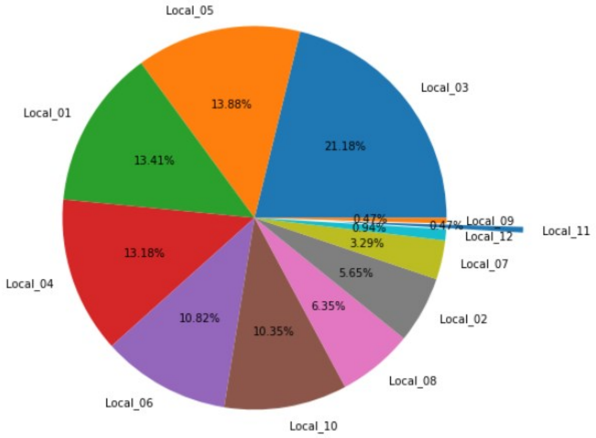


Figure 7: Accidents reported by Local – Percentages

3.1.4.3 Gender

Table 5: 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.

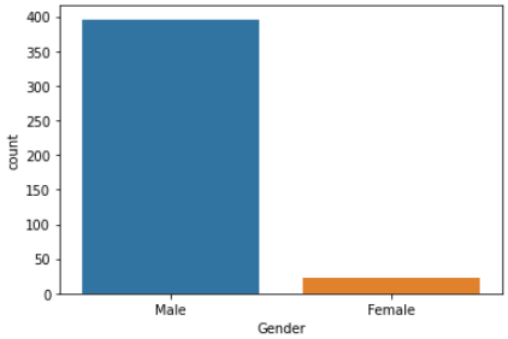


Figure 8: Accidents reported based on Gender of Person injured – Counts

3.1.4.4 Industry Sector

Table 6: Accidents reported by Industry Sector

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

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.

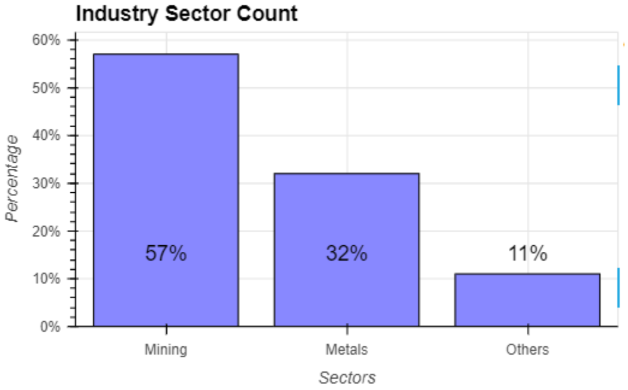


Figure 9: Accidents reported by Industry Sector - Percentages of Counts

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 7: 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 8: 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%

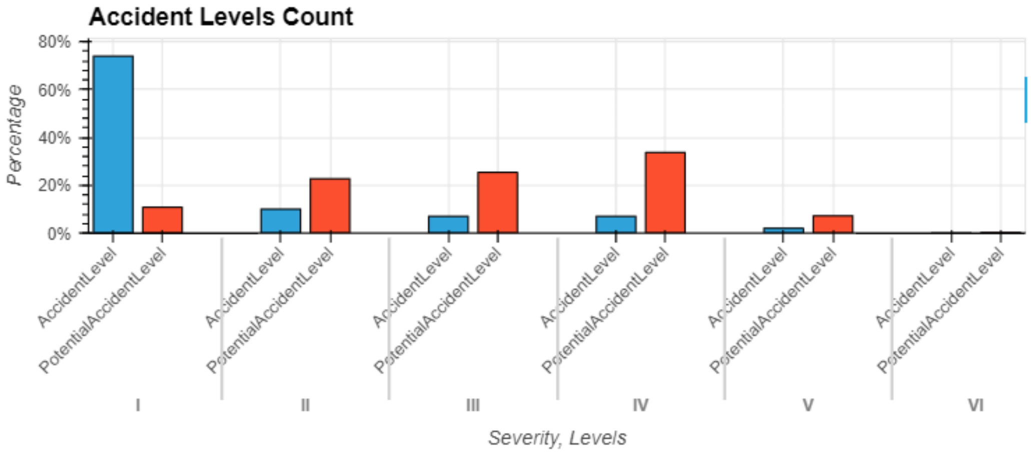


Figure 10: Accidents reported by Accident Level and Potential Accident Level – Percentages

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 punishments. 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 9: Accidents reported based on Employee Type of Person injured

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

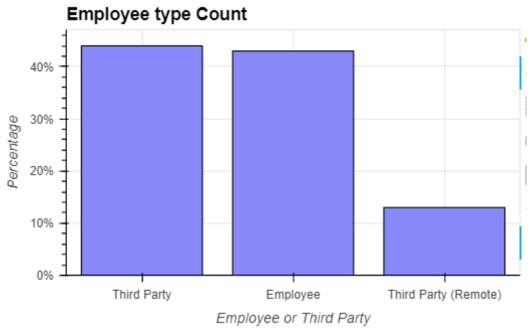


Figure 11: Accidents reported based on Employe Type of Person injured – Percentages

Number of Accidents involving Employee and Third Party are 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 clear 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 10: 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%

Critical Risk	Count	%
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%
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”.

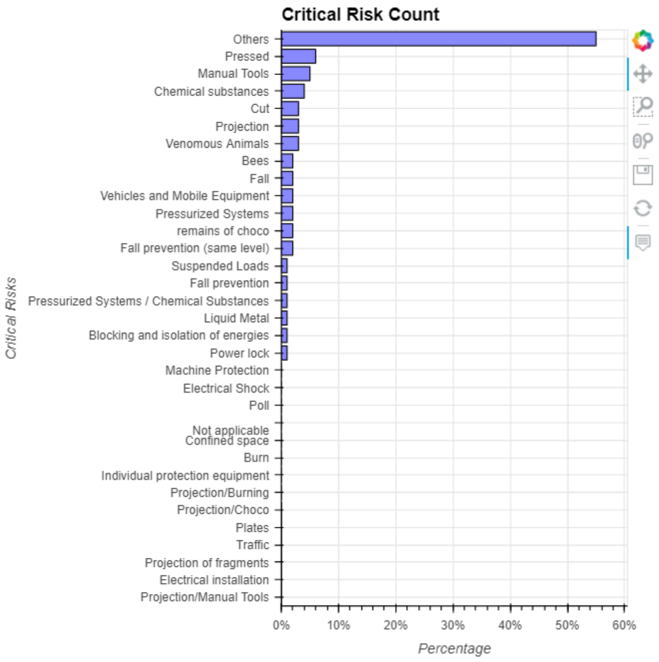


Figure 12: Accidents reported by Critical Risk – Percentages

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. For, 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.

3.1.5 Multivariate Analysis

3.1.6 Hypothesis Testing

3.1.7 Data Analysis Summary

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 11: 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 user might use in the text input	from nltk.stem import WordNetLemmatizer lemmatize method	The word tokens are lemmatized
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		

Pre-Processing Step	Purpose	Method Used	Output
	consistency and simplifying training		
Ignore Numeric Text	Ignore numeric text and pick up only alphabetic text		

Example:

> **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"

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

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

3.2.1.1 Length of Description (Characters)

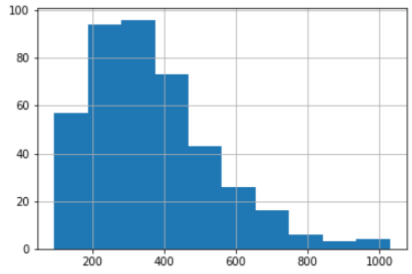


Figure 13: Histogram of Length of Description before NLP Pre-Processing

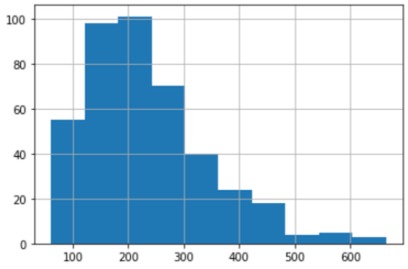


Figure 14: Histogram of Length of Description after NLP Pre-Processing

3.2.1.2 Word Count in Description

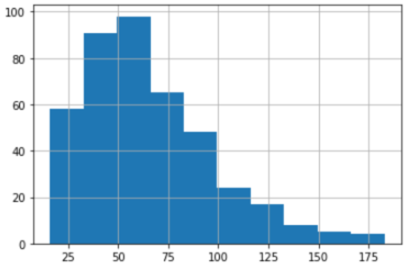


Figure 15: Histogram of Number of Words in Description before NLP Pre-Processing

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

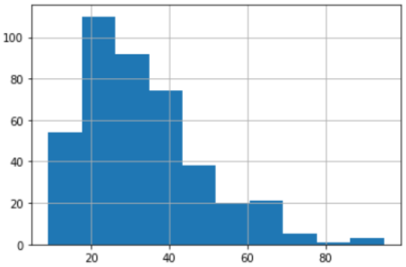


Figure 16: Histogram of Number of Words in Description after NLP Pre-Processing

3.2.1.3 Average Length of Words

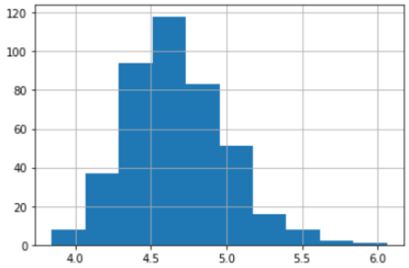


Figure 17: Average Word Length in Description before NLP Pre-Processing

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

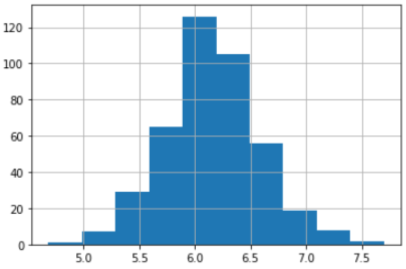


Figure 18: Average Word Length in Description after NLP Pre-Processing

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:

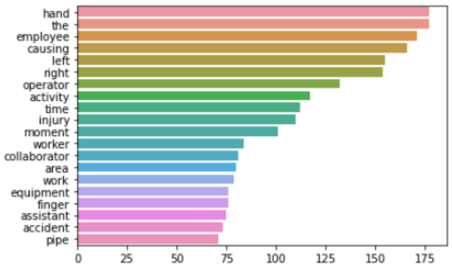


Figure 19: Frequency of Word Sequences N-Gram (N=1)

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.

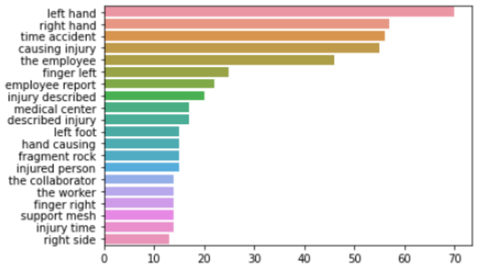


Figure 20: Frequency of Word Sequences N-Gram (N=2)

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.

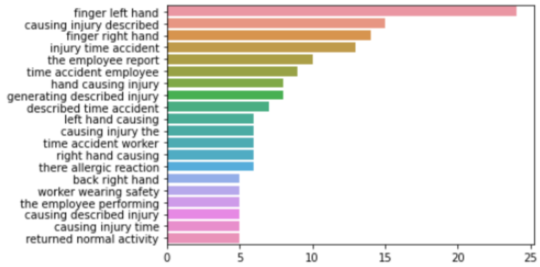


Figure 21: Frequency of Word Sequences N-Gram (N=3)

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.

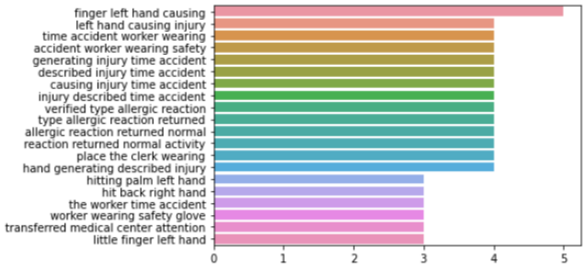


Figure 22: Frequency of Word Sequences N-Gram (N=4)

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

3.2.2 Featurization, Model Selection & Tuning Strategy

3.2.3 Chatbot Architecture Evaluation

3.2.4 User Interface and Interaction Evaluation

3.2.5 Deployment Architecture Evaluation

4 MODEL TRAINING AND TESTING

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

Table 12: Model Selection - Comparison of Metrics

Model					

5 CHATBOT TRAINING AND TESTING

Rasa Chatbot

Custom developed Python based 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