greatlearning

CAPSTONE PROJECT

NLP CHATBOT

The Team:

Kajal Jaiswal

Damini Tiwari

Ashwini Kumar

Anis Uddin

Mentor:

Sravan Malla

Presentation Overview

- Problem Introduction
- Exploratory Data Analysis
- Data Preprocessing
- Model Building
- Chatbot Deployment

Summary of Problem Statement

Our AIML Capstone project is to design a ML/DL based chatbot utility which can help industry professionals to highlight the safety risks as per the incident description reported.

The database comes from one of the biggest industries in Brazil and in the world. It is an urgent need for industries/companies around the globe to understand why employees still suffer some injuries/accidents in plants.

Goals

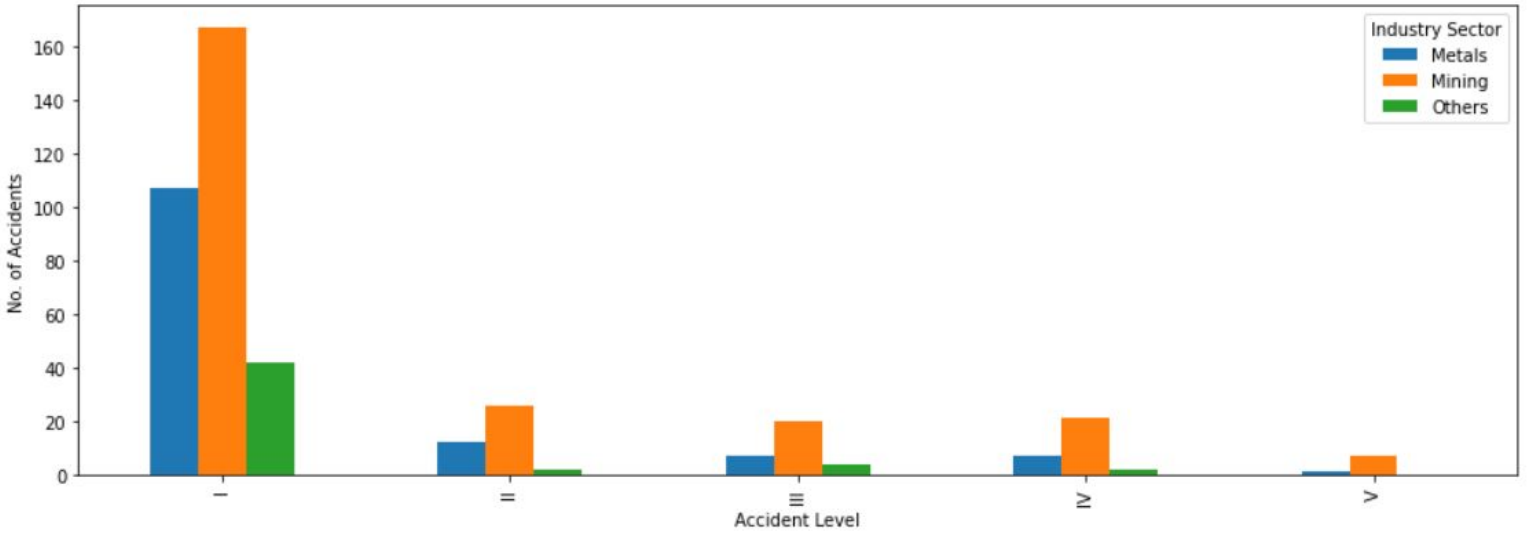
1. Designing an AIML chatbot that recognizes the Accident Level, Potential Accident Level, Critical Risks involved with the accident based on the descriptions provided.

2. Designing and deploying a clickable UI based chatbot interface.

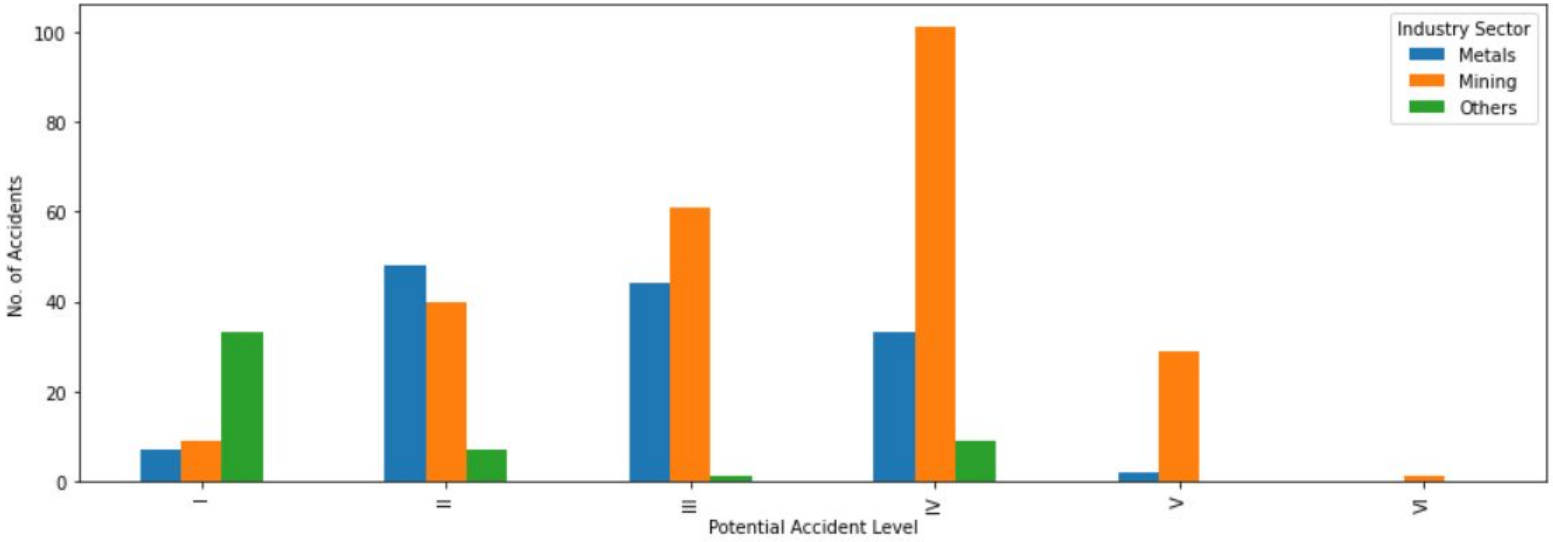
The Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
		Data	Countries	Local	Industry Sector	Accident Level	Potential Accide	Genre	Employee or Thir	Critical Risk	Description			
0		2016-01-01 0:00:	Country_01	Local_01	Mining	I	IV	Male	Third Party	Pressed	While removing the drill rod of the Jumbo 08 for maintenance, the supe			
1		2016-01-02 0:00:	Country_02	Local_02	Mining	I	IV	Male	Employee	Pressurized Sys	During the activation of a sodium sulphide pump, the piping was uncou			
2		2016-01-06 0:00:	Country_01	Local_03	Mining	I	III	Male	Third Party (Rem	Manual Tools	In the sub-station MILPO located at level +170 when the collaborator w			
3		2016-01-08 0:00:	Country_01	Local_04	Mining	I	I	Male	Third Party	Others	Being 9:45 am. approximately in the Nv. 1880 CX-695 OB7, the person			
4		2016-01-10 0:00:	Country_01	Local_04	Mining	IV	IV	Male	Third Party	Others	Approximately at 11:45 a.m. in circumstances that the mechanics Ant			
5		2016-01-12 0:00:	Country_02	Local_05	Metals	I	III	Male	Third Party (Rem	Pressurized Sys	During the unloading operation of the ustulado Bag there was a need to			
6		2016-01-16 0:00:	Country_02	Local_05	Metals	I	III	Male	Employee	Fall prevention (s	The collaborator reports that he was on street 09 holding in his left han			
7		2016-01-17 0:00:	Country_01	Local_04	Mining	I	III	Male	Third Party	Pressed	At approximately 04:50 p.m., when the mechanic technician José of th			
8		2016-01-19 0:00:	Country_02	Local_02	Mining	I	IV	Male	Third Party (Rem	Others	Employee was sitting in the resting area at level 326 (raise bore), wher			
9		2016-01-26 0:00:	Country_01	Local_06	Metals	I	II	Male	Third Party	Chemical substa	At the moment the forklift operator went to manipulate big bag of bioxic			
10		2016-01-28 0:00:	Country_01	Local_03	Mining	I	III	Male	Employee	Others	While installing a segment of the polyurethane pulley protective lyner -			

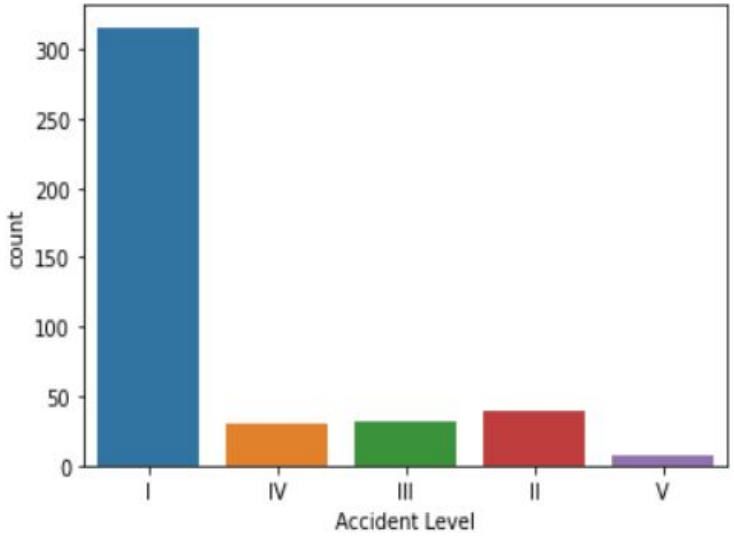
‘Accident’ count vs ‘Accident Level’

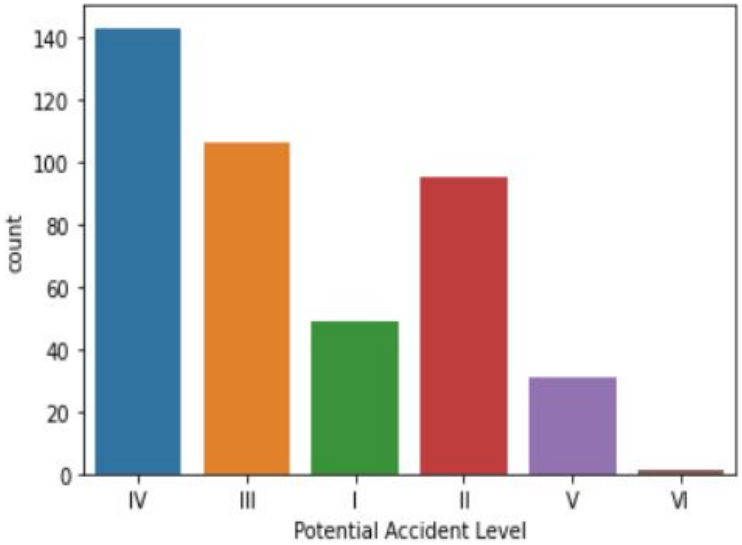


‘Accident’ count vs ‘Potential Accident Level’



Count Plots for ‘Accident Level’ and ‘Potential Accident Level’





Word Cloud

The ‘Descriptions’ feature depicted as a word cloud.

Data Preprocessing

1. Eliminate any special characters
2. Remove Stopwords
3. Remove all punctuation
4. Lemmatization
5. Eliminating the 20 most rare words
6. Spellcheck
7. Tokenization
8. Sequence Padding

Before Preprocessing	After Preprocessing
data.Description[1] During the activation of a sodium sulphide pump, the piping was uncoupled and the sulfide solution was designed in the area to reach the maid. Immediately she made use of the emergency shower and was directed to the ambulatory doctor and later to the hospital. Note: of sulphide solution = 48 grams / liter.	data.Description[1] activation sodium sulphide pump piping uncoupled sulfide solution designed area reach maid immediately made use emergency shower directed ambulatory doctor later hospital note sulphide solution gram liter
data.Description[2] In the sub-station MILPO located at level +170 when the collaborator was doing the excavation work with a pick (hand tool), hitting a rock with the flat part of the beak, it bounces off hitting the steel tip of the safety shoe and then the metatarsal area of the left foot of the collaborator causing the injury.	data.Description[2] sub station milo located level collaborator excavation work pick hand tool hitting rock part beak bounce hitting steel tip safety shoe metatarsal area left foot collaborator causing injury

Model Building

Below are the models that were used:

1. Logistic Regression
2. Random Forest
3. Support Vector Machine (SVM)
4. Ensemble Techniques : Boosting & Bagging
5. Neural Network
6. LSTM

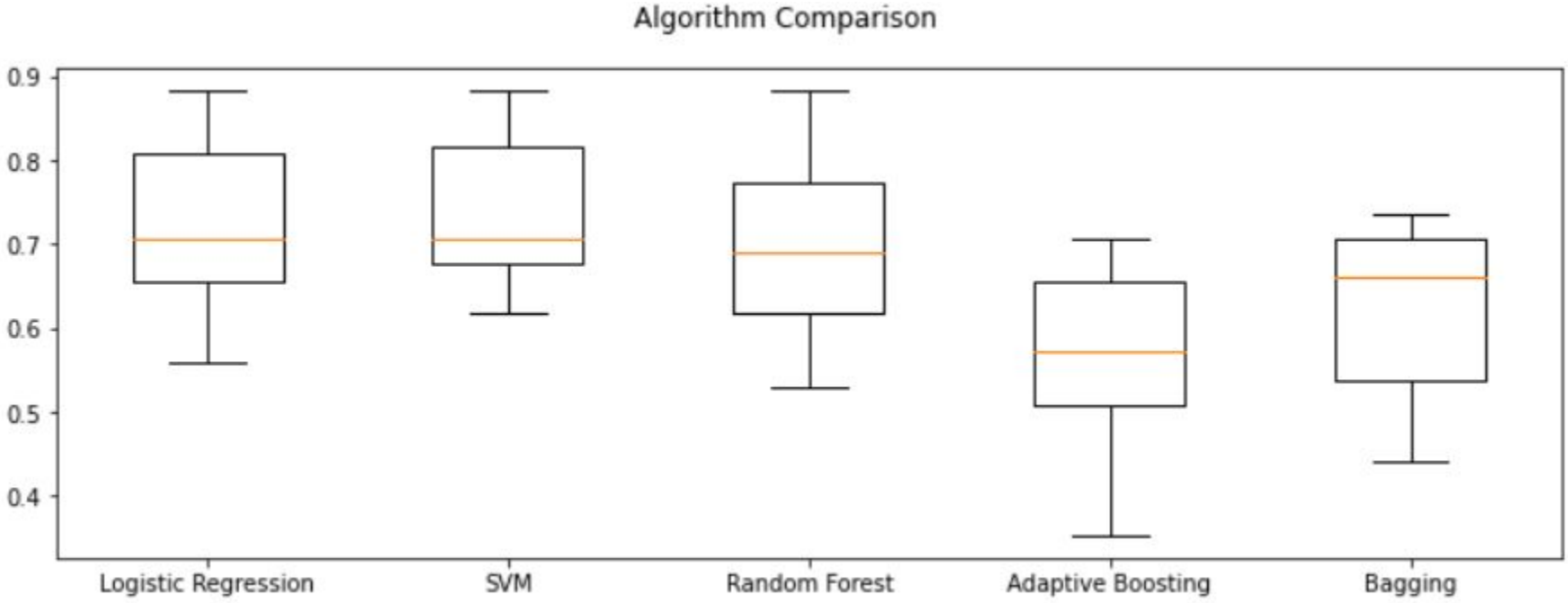
Training Metrics Accuracy Of AL

	Model	accuracy	F1 score	precision	recall	roc_auc
1	Logistic Regression	0.782353	0.805882	0.68827	0.805882	0.878676
2	Random Forest	0.926471	0.951662	0.921036	0.926471	0.960662
3	SVM	0.911765	0.913108	0.851414	0.911765	0.945221
5	Bagging	0.988235	0.99115	0.984741	0.988235	0.993382
4	Boosting	0.782353	0.837209	0.731602	0.847059	0.901471
6	NN	0.726471	0.678039	0.635662	0.726471	0.902117
7	LSTM	0.779412	0.761332	0.852957	0.779412	0.958946

Testing Metrics Accuracy Of AL

	Model	accuracy	F1 score	precision	recall	roc_auc
1	Logistic Regression	0.782353	0.805882	0.68827	0.805882	0.878676
2	Random Forest	0.926471	0.951662	0.921036	0.926471	0.960662
3	SVM	0.911765	0.913108	0.851414	0.911765	0.945221
5	Bagging	0.988235	0.99115	0.984741	0.988235	0.993382
4	Boosting	0.782353	0.837209	0.731602	0.847059	0.901471
6	NN	0.726471	0.678039	0.635662	0.726471	0.902117
7	LSTM	0.779412	0.761332	0.852957	0.779412	0.958946

Cross Validation and ML models accuracy projection as box plot.



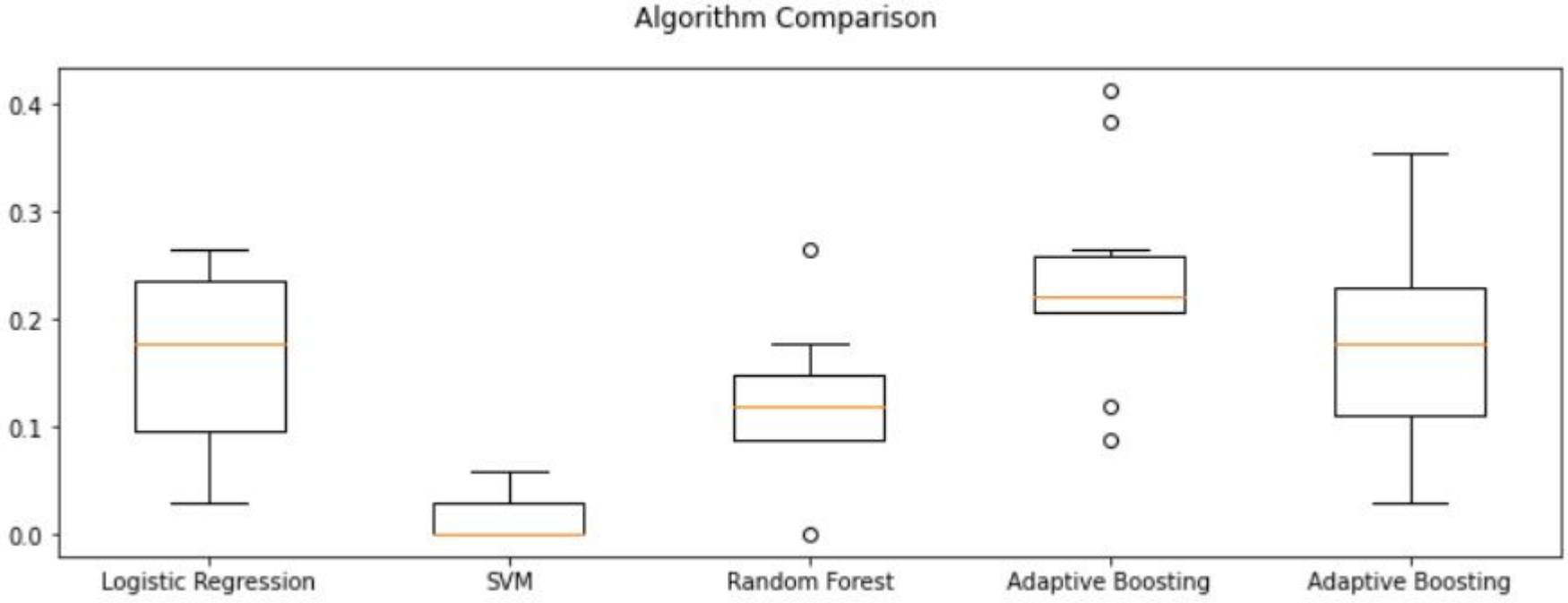
Training Metrics Accuracy Of PAL

	Model	accuracy	F1 score	precision	recall	roc_auc
1	Logistic Regression	0.364706	0.500935	0.4223	0.394118	0.667157
2	Random Forest	0.905882	0.949153	0.92648	0.905882	0.952451
3	SVM	0.591176	0.743068	0.693382	0.591176	0.795588
5	Bagging	0.994118	0.99705	0.995588	0.994118	0.997059
4	Boosting	0.441176	0.594697	0.520181	0.461765	0.715686
6	NN	0.241176	0.168514	0.129499	0.241176	0.600702
7	LSTM	0.826471	0.837108	0.856888	0.826471	0.961672

Test Metrics Accuracy Of PAL

	Model	accuracy	F1 score	precision	recall	roc_auc
1	Logistic Regression	0.141176	0.2	0.255948	0.152941	0.513725
2	Random Forest	0.211765	0.318584	0.333193	0.211765	0.586275
3	SVM	0.0823529	0.225806	0.267949	0.164706	0.533333
5	Bagging	0.994118	0.99705	0.995588	0.994118	0.997059
4	Boosting	0.235294	0.353846	0.320654	0.270588	0.592157
6	NN	0.164706	0.261339	0.529563	0.176471	0.539571
7	LSTM	0.564706	0.610457	0.686957	0.564706	0.776458

Cross Validation and ML models accuracy projection as box plot.

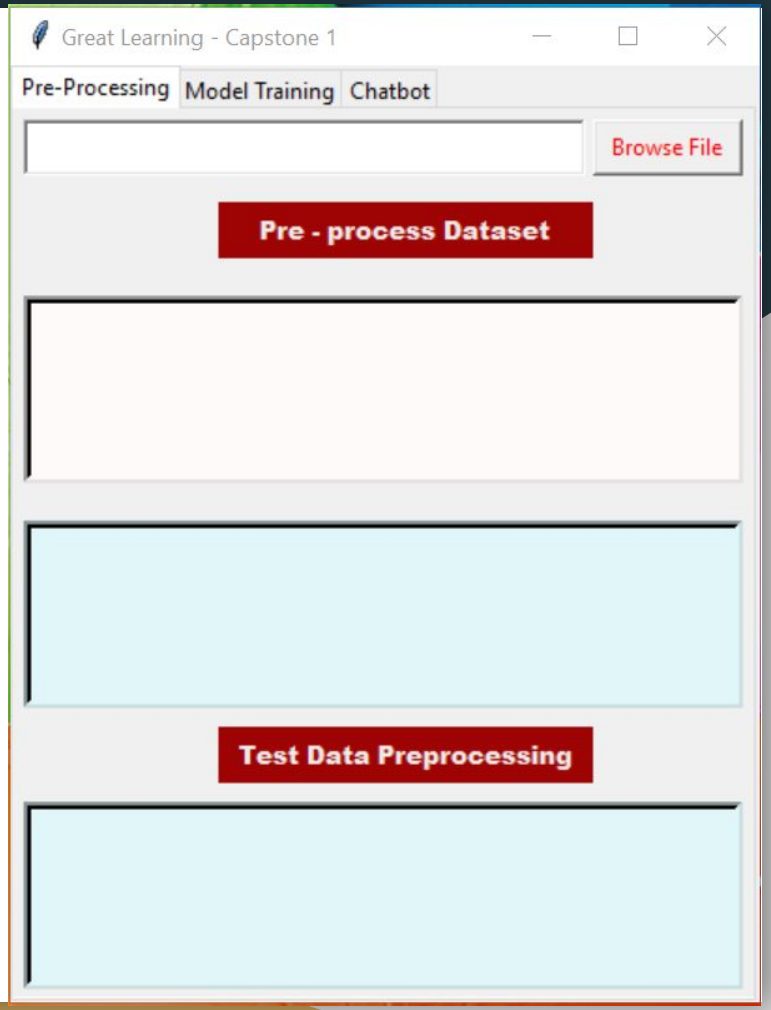


Model Selection

From the obtained results, we have selected the LSTM Model for this classification problem as it has shown all round performance in both test and train accuracies. The ML models have been observed to be performing great in train data and fairly poor in test data.

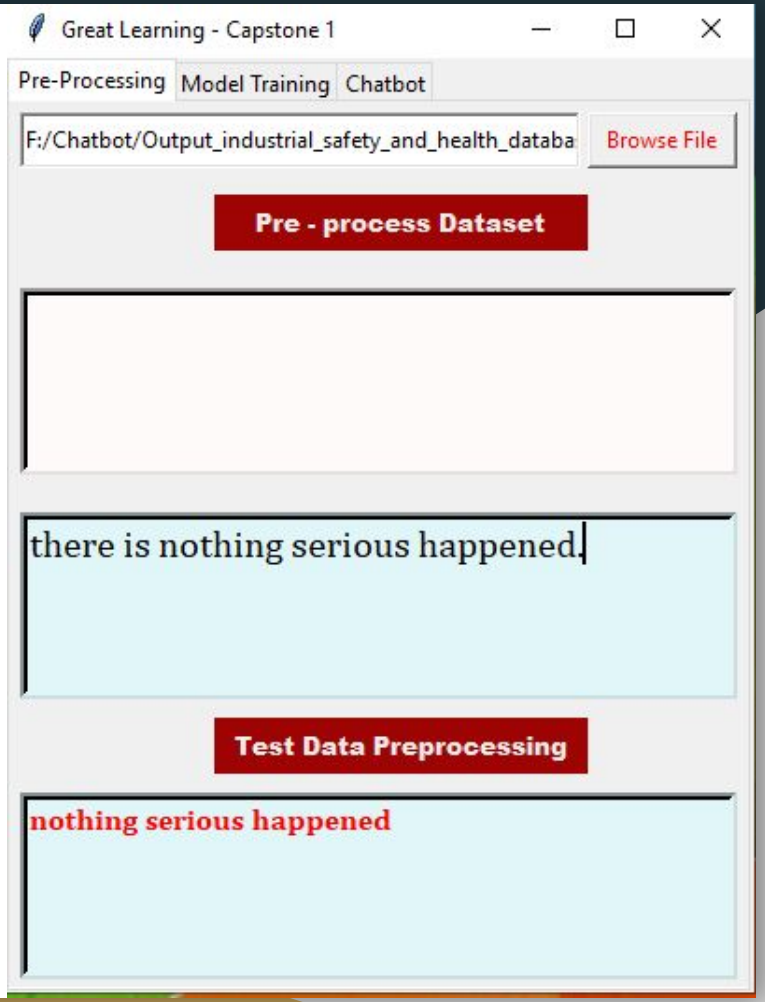
ChatBot Deployment

This is the Main Screen of the Chatbot UI along with the 3 buttons shown as tabs.

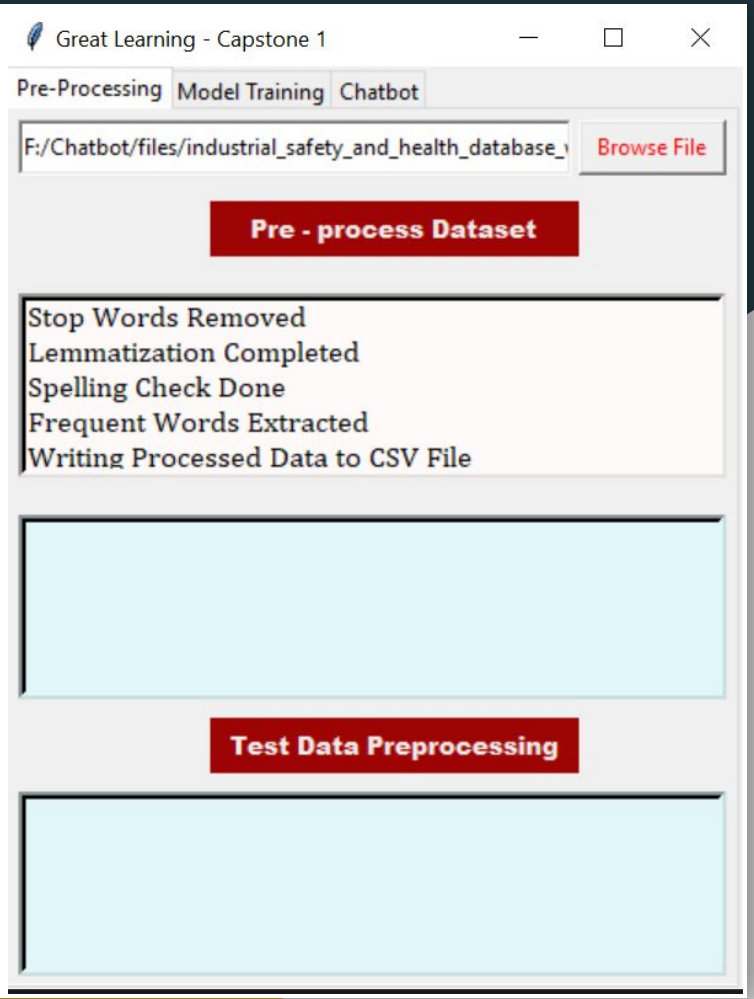


Initial Preprocessing

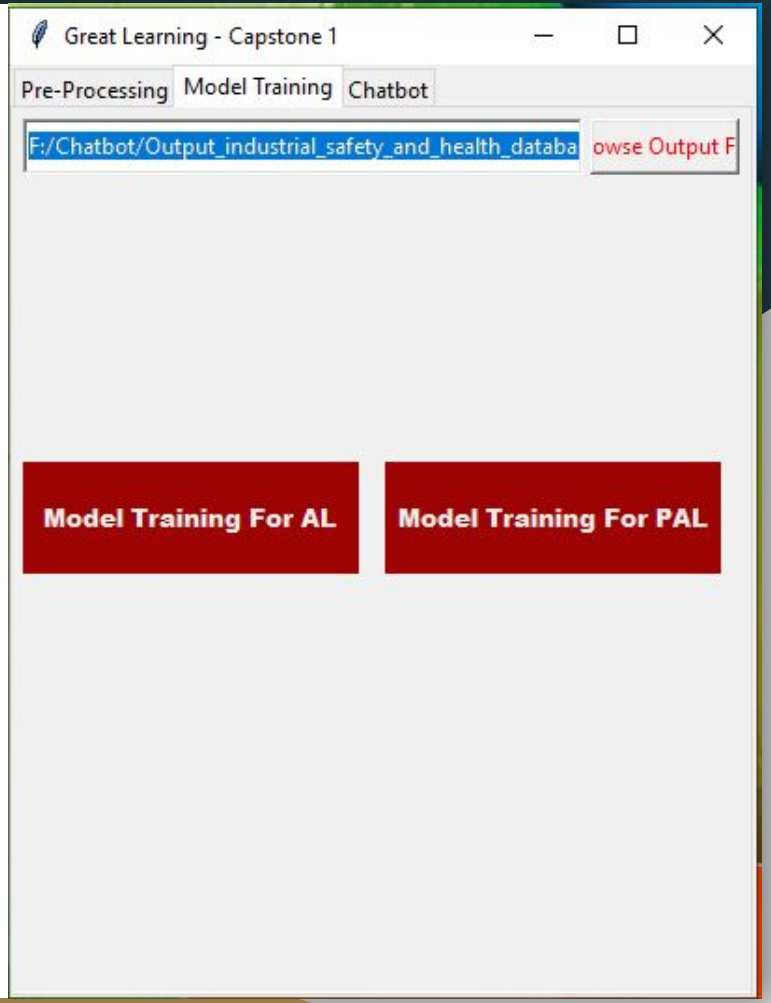
Here we are loading the dataset into the chatbot and running the preprocessing function to prepare an output file that can be used for making predictions later.

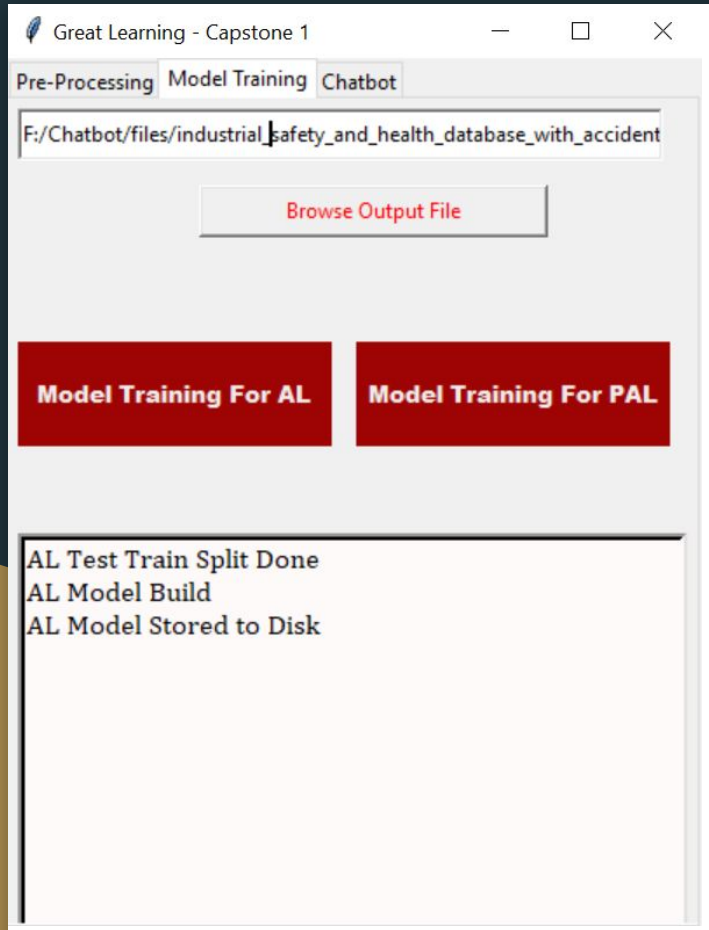


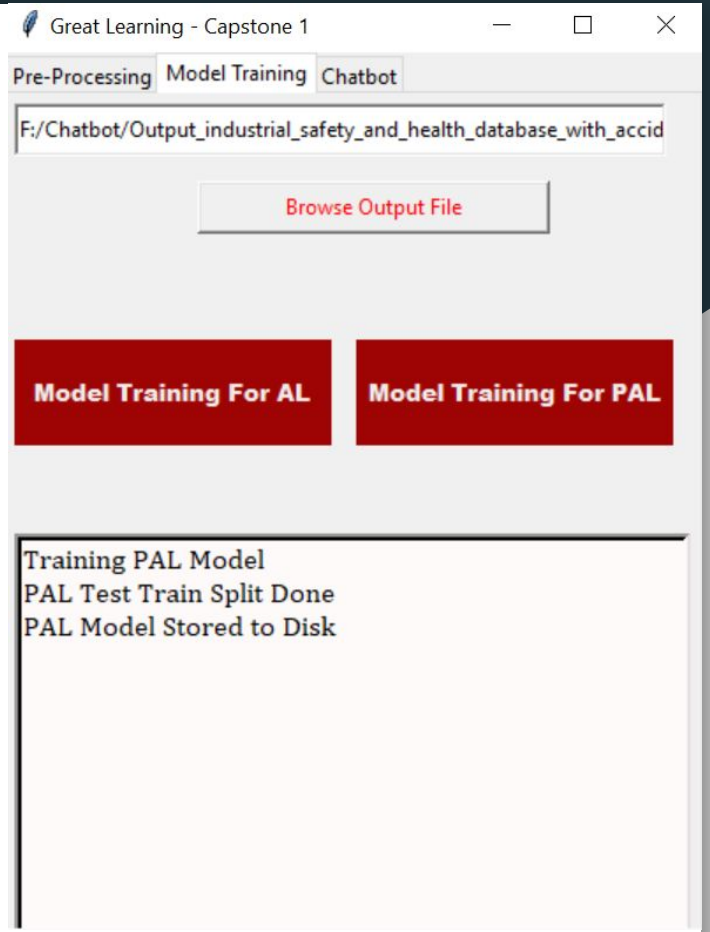
Here we are running the Preprocessing on the entire Dataset. It is showing the various functions that are being run in the background like removing StopWords, Lemmatization, Spell Check etc. The output is then saved as a CSV file



Model Training Button

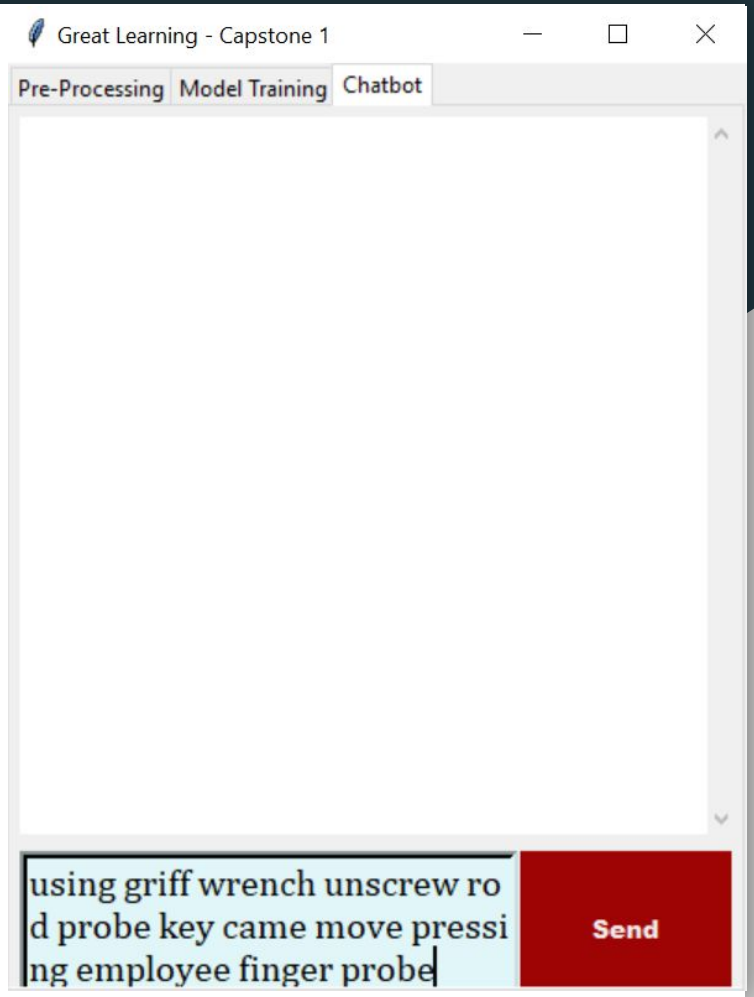


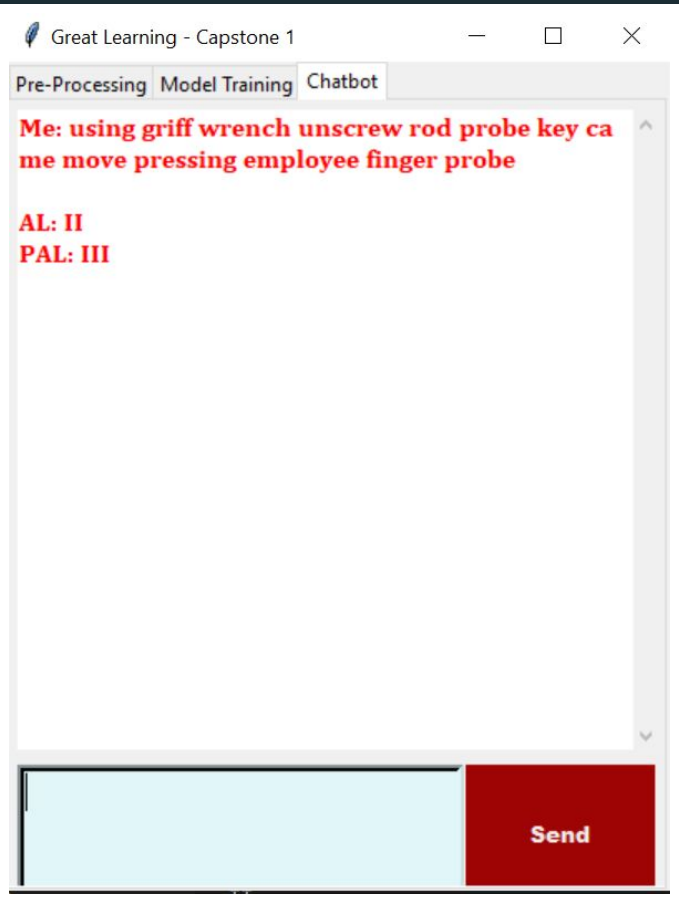


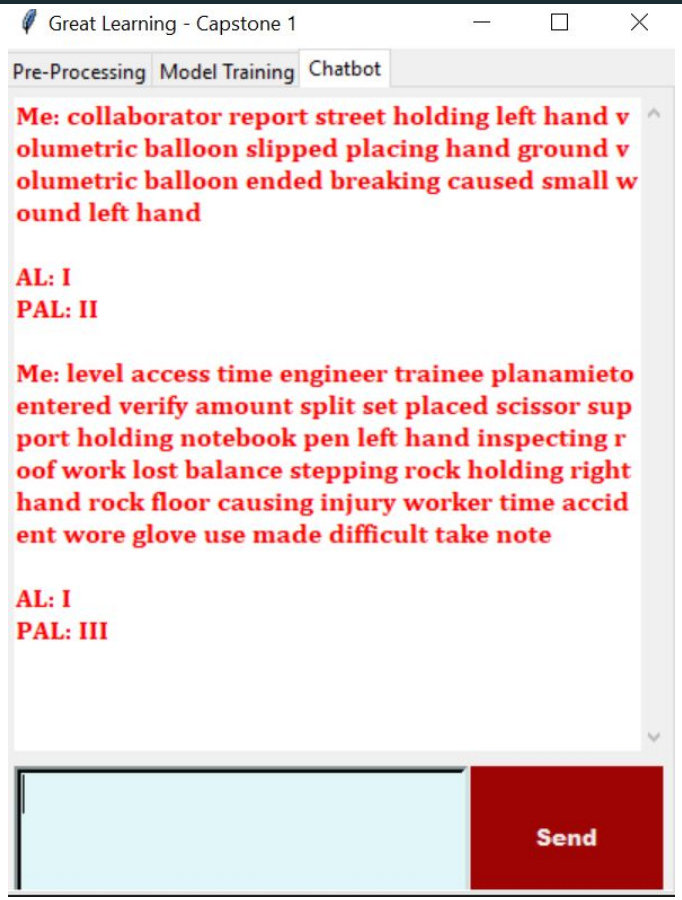


We can see the chatbot is performing the training on AL and PAL as well as splitting the data set and storing the model locally.

Finally, we are now typing out a description in the Chatbot button tab. This will be the input data for our models.







Here is the output of the chatbot where it is predicting the Accident level and Potential Accident Level based on the description of the accident.

Conclusion

We can conclude that we have successfully built a chatbot interface that can analyze a given text input and predict the required outcomes based on this.

Next steps of improvement to the project is discussed in the next slide.

Next Steps

Here are few points that were discussed within the team:

1. The chatbot can be made to be more human-like with some opening statements like ‘Welcome! I am your accident assistant. Please describe in details what has happened and I will help you to analyze the Accident Level as well as the Potential Accident Level’ and intermittent statements like ‘Please give me a moment, I am analyzing’ etc
2. There would need to be rules added to the Chatbot such as minimum length of input to be accepted so that a simple ‘Hi’ or ‘Hello’ is not considered as input. This would also eliminate poor predictions as a certain length of description is required for accuracy. In the case where this rule is not met, we can have an error statement such as ‘Sorry, I would need more detail to help you out. Please elaborate’
3. Hosting the GUI over a web based service for easier access as well as faster interactions as the training and testing data can be stored there itself and there wouldn’t be a need to run it over locally everytime.

Thank you

greatlearning

On behalf of the team, we wholeheartedly thank you for this wonderful opportunity to learn, experiment, implement and grow.