

# FINAL PROJECT REPORT

---

Team HDPZ



Project Title	SmartAlert: Optimizing Emergency Response Time
Date Started	01-09-2023
Date Completed	05-01-2023
Project Sponsor	Dr. Anna Farzindar ( <i>USC</i> ) & Dr. Alex Liu ( <i>RMDS Lab</i> )

## Team Details

**Haoran Shu**

shuhr@usc.edu

**Praveen Allu**

pallu@usc.edu

**Duyen Nguyen**

duyenngu@usc.edu

**Ziyan Ping**

ziyanpin@usc.edu

# TABLE OF CONTENTS

<b>Chapter 1.....</b>	<b>1</b>
Executive Summary.....	1
<b>Chapter 2.....</b>	<b>2</b>
Project Objectives.....	2
<b>Chapter 3.....</b>	<b>3</b>
Lean Six Sigma Project.....	3
3.1 Define Phase.....	3
3.1.1 Customer Satisfaction.....	3
3.1.2 Tools Application.....	4
3.2 Measure Phase.....	5
3.2.1 Process Mapping.....	5
3.2.2 The Vital Few.....	5
3.2.3 Data Exploration and Preparation.....	6
3.2.4 Tools Application.....	8
3.3 Analysis Phase.....	9
3.3.1 Selecting charts for Analysis.....	9
3.3.2 Value Function.....	9
3.3.3 Sources of Variation.....	9
3.3.4 Potential Solutions.....	11
3.3.5 Tools Application.....	12
3.4 Improve Phase.....	13
3.4.1 Solution Evaluation.....	13
3.4.2 Recommended Solution.....	17
3.4.3 Pilot Design.....	17
3.4.4 Work Breakdown Structure.....	18
3.5 Control Phase.....	20
3.5.1 Control Solutions Considered.....	20
3.5.2 Control Solution Implemented.....	20
3.6 Result and System Implementation.....	22
3.6.1 Machine Learning Approaches.....	22
3.6.2 System Implementation.....	23
3.6.3 Prototype and Demonstration.....	25
<b>References.....</b>	<b>28</b>
<b>Appendix.....</b>	<b>29</b>

## Chapter 1

# Executive Summary

- Natural disasters can occur unexpectedly and can cause significant damages to people's lives and infrastructures. Quick response times are critical for natural disaster management in a smart city as it can play a vital role in mitigating the effects of natural disasters.
- The current system for responding to emergencies can often be slow and inefficient in assessing the severity of a situation and providing timely assistance. This can result in valuable time being lost and the situation deteriorating, leading to increased damage and potential loss of life.
- Lack of timely and accurate information during a disaster can lead to chaos, confusion, and delay response times, putting lives at risk.
- HDPZ, an tech consulting team, aims to find the solution to the mentioned problem by developing a system that monitors social media platforms, such as Twitter, using machine learning algorithms to identify and categorize relevant tweets into disaster types, severity levels and locations, improving response times during natural disasters.
- Our end product is an web application that notifies the disaster management departments with real-time information about the situation on the ground, shows alerts on a map to local residents and the governments, and lastly, agents have the full control of verifying and updating these unexpected situations.

## Chapter 2

# Project Objectives

- Utilize historical tweet datasets to analyze and classify spam and disaster-related tweets.
- Disaster-related tweets will then be used to classify into potential disaster type.
- Disaster-related tweets will also be used for location extraction using Spacy.
- Implement scraper to extract tweets and analyze them in real-time to improve response times during natural disasters.
- Create a web application for agents to receive first alerts about potential natural disasters, the disaster types and locations.
- Develop a user-friendly web application capable of monitoring social media platforms, such as Twitter, to provide precise information to disaster management departments, governments, and the general public.

## Chapter 3

# Lean Six Sigma Project

The project utilizes the DMAIC technique, which consists of five stages, namely define, measure, analyze, implement, and control. The report also includes a sixth section that provides technical information on the machine learning algorithms used, as well as the system implementation, prototype, and demonstration. Details will be shown below.

## 3.1 Define Phase

In Lean Six Sigma methodology, the initial stage is referred to as define. In this phase, our team worked on a Project Charter, developed a high-level process maps, as well as getting voice of customers as the requirements for the process.

### 3.1.1 Customer Satisfaction

Victims of emergency incident:

- We talked to two disaster victims to know their thoughts and experiences about how the communication and assistance they recieved were during the crisis.
- The first thing we learned is that both interviewees agreed that quick response time is super important in emergency situations as it can help alleviate the stress from the situation.
- Both interviewees suggested that there is a lot of room for improvement in the communication between victims and emergency services, and in the assistance they received during their respective crises.
- Both users think there should be some improvements in the emergency reporting and notification system, as well as better traffic management to make emergency response times faster.
- Additionally, they would like to have a system where they can both report incidents and receive emergency notifications in their vicinity.

Disaster departments liaisons:

- During an interview, the user expressed that having a system for receiving emergency notifications quickly is essential for timely assistance, especially in situations where the victim's life is at risk. Such a system could provide real-time updates on the situation and ensure that emergency services can respond as soon as possible.
- Furthermore, the user suggested adding a feature that predicts the severity of the emergency in the system. This feature would enable the dispatch team to allocate personnel and vehicles accordingly, which would optimize the use of resources and reduce response times.
- The user highlights the importance of prompt and efficient communication during emergency situations. The user believes that a reliable emergency notification system, combined with a feature that predicts the severity of the emergency, would improve the overall emergency response and save more lives.

### **3.1.2 Tools Application**

Among the tools used in the define phase, the project charter was found to be the most useful in creating a concise and comprehensive overview of all strategic factors related to the project. This aided in better communication with stakeholders and improved feedback collection. Early input gathering, particularly in the initial stages when the project is beset with uncertainty, is important to prevent misalignment issues later on. By providing clear delineation of responsibilities, constraints, and milestones, the project can be set on a solid foundation for success.

## 3.2 Measure Phase

Measurement takes two to three weeks, depending on the project inputs. To obtain high-quality data, collaboration with all pertinent data providers in particular is essential. Establishing a baseline for the current process, gathering data, validating the measurement system, and figuring out the capabilities of the process are the main goals of the measure phase.

### 3.2.1 Process Mapping

To illustrate the workflow for this project, several process maps were made. There are several different process maps that can be used, including a SIPOC, High-level Process Map, Common Process Map, Detailed Process Map, and Functional Process Map. Making these maps was essential to learning more about the inputs and outputs that each workflow phase is accountable for.

This was critical in establishing a methodology for combining the results of three models developed into a single, user-friendly product. The work done on the SIPOC was used to create a task list, which is detailed in the Detailed Process Map. This phase served as evaluation points for particular project tasks and allowed the team to stick to a defined continuous development strategy.

### 3.2.2 The Vital Few

We recognize the importance of developing an effective disaster response system, and we believe that the following three components are vital to its success. Firstly, our disaster-related tweet classification model must be accurate in identifying and categorizing tweets into different disaster types, severity levels, and locations. This model plays a crucial role in ensuring that we are notified of relevant tweets in real-time, which will enable us to respond quickly and efficiently. Secondly, the location extraction model, powered by Spacy, must be precise in identifying the locations mentioned in the tweets. This will enable us to display alerts on a map and notify the relevant disaster management departments immediately. Lastly, the real-time scraper for Twitter is the backbone of the entire system and needs to be reliable and efficient. This scraper is responsible for extracting tweets in real-time and analyzing them to improve response times during natural disasters. As a team, we recognize that these three components are the most vital aspects of our disaster response system, and we are committed to ensuring their accuracy and effectiveness.

### 3.2.3 Data Exploration and Preparation

We obtained our data for model development from <https://crisisnlp.qcri.org>, specifically, [https://crisisnlp.qcri.org/humaid\\_dataset](https://crisisnlp.qcri.org/humaid_dataset) (Event-wise dataset (set1)) and <https://crisisnlp.qcri.org/> (resource12).

These 2 Twitter datasets consist of more than 50 000 manually annotated tweets collected during 19 major natural disaster events, including earthquakes, hurricanes, wildfires, and floods, that occurred around the world between 2016 and 2019. Their data attributes are shown below.

Eyewitness dataset - <https://crisisnlp.qcri.org/> (resource12)

<b>Attribute</b>	<b>Description</b>
Category	Categorical attribute (eyewitness, non-eyewitness, unknown)
Tweet text	The textual portion of a tweet
Disaster	Categorical attribute (flood, earthquake, hurricane, and wildfire)

Table 1. Eyewitness dataset's attributes

Events set - [https://crisisnlp.qcri.org/humaid\\_dataset](https://crisisnlp.qcri.org/humaid_dataset) (Event-wise dataset (set1)) - each event is a disaster with train, dev, and test datasets.

<b>Attribute</b>	<b>Description</b>
Class label	Categorical attributes (annotations consisting of 11 categories)
Tweet text	The textual portion of a tweet

Table 2. Eventsset's attributes

We preprocessed tweet text using several techniques below:

- Expand Contractions: This method will be used to expand the contractions, such as "I'm" to "I am" in the text. After this, emergency tweets will become easier to read and



---

comprehend.

- Lower Case: Making the text lowercase reduces the complexity of the data and improves the efficiency of the analysis. It also makes sure that similar-sounding terms are not treated differently.
- Remove Punctuation: This method eliminates all punctuation from the text since it adds more dimensions to the data without adding any significant information.
- Remove all non-alphabetic characters: With this method, all non-alphabetic characters will be removed, like numerals and special characters, and the data is prepared for the next cleaning step. This improves the text's coherence and makes it simpler to do the analysis.
- Stopwords are eliminated: Since stopwords like "the," "a," and "an" don't convey any useful information, this strategy eliminates all of them. This makes the analysis more effective by lowering the data's dimensionality.
- Stemming and lemmatization: These techniques are used to reduce the words to their base form, thus reducing the dimensionality of the data and making the analysis more efficient.
- Remove white spaces: This technique removes all the extra white spaces present in the text, as they do not provide any meaningful information and only increase the dimensionality of the data. NLP packages the **nlTK** library to remove white spaces and stop words.
- Removing URLs: Removing URLs is used to remove any URLs present in the text, like <https://twitter.com>, as they do not provide any meaningful information and only increase the dimensionality of the data.
- Removing HTML tags: This approach is used to get rid of any HTML tags that are present in the text since they add more dimensions to the data without adding any significant information.
- Removing emoji and special characters: This technique was used to remove any emojis or special characters present in the text, like special letters `$%$^#@`, as they do not provide any meaningful information and only increase the dimensionality of the data.

### **3.2.4 Tools Application**

In case of the training and evaluation phase of the project the data was collected from crisisNLP. We used Apify twitter scraper to extract the tweets. Using this API provided a simple and manageable approach of getting all of the necessary data for inference phase of our project. Using these services to acquire information proved to be pretty advantageous and saved us a lot of time in the early stages of the project because we didn't have to commit time to exploring many alternative sources for data collection. As a result, our team was able to shift its emphasis to other stages of the project, such as model creation and assessment. However, as Twitter hid its search functionality behind login on 04/21/2023, the latest inference data we could collect was up to 04/20/2023.

Process mapping tools, including SIPOC 8, common 9, detailed 10, functional 11, were valuable for their ability to clearly define the inputs and outputs for each step of the project. This was crucial in breaking down the entire pipeline into simpler modules that could be easily and quickly controlled.

## 3.3 Analysis Phase

During the analyze step, the process map is evaluated to improve efficiency, identify possible root causes, and determine key factors or inputs that have a significant impact on the output.

### 3.3.1 Selecting charts for Analysis

By utilizing the DMAIC method, at the analyze phase, we first used the fishbone to summarize all the causes that would lead to poor model performance. And then we used Pareto's Chart to further illustrate the reasons listed above in more detail. This chart also helped us identify potential causes that had a significant impact on our project's performance. Finally, by using 5 whys, we could get to the heart of an issue and identify the underlying cause.

### 3.3.2 Value Function

In order to perform root cause analysis of potential problems in a project, it is crucial to understand how the project creates value for its customers. Value creation can be viewed as the process of converting inputs into desired outputs through a particular function. Therefore, it is important to clearly define the project's value function before conducting root cause analysis. This will provide a framework for understanding the causes of any problems that arise and help to identify solutions that will improve the project's ability to deliver value to its customers.

Our project focused on detecting and classifying tweets indicating potential natural disasters. We were able to classify relevant tweets, their disaster type and locations. Customers can access our final product through an interactive web application.

### 3.3.3 Sources of Variation

Main sources of variation lie in two areas:

- Machine Learning Models:
  - Classification models: Random forest, Support vector machine, and Feedforward neural networks were taken into consideration for classification model. Random forest is known for its high accuracy and ability to handle large datasets with high-dimensional feature spaces. It also performs well with noisy data and missing values. However, it can be slower in training and requires more computational resources than other algorithms. SVM, on the other hand, is effective in handling high-dimensional data and performs well with a small

number of training samples. It also has good generalization performance. FFNN is a powerful algorithm for classification tasks due to its ability to capture complex patterns in data. It can handle both continuous and categorical data and has good generalization performance. However, FFNN requires a large amount of training data and can be computationally expensive. It is also prone to overfitting if not properly regularized.

- Topic modeling: LDA, PLSA, and BERTopic are all popular topic modeling techniques with their own advantages. LDA is a well-established probabilistic model that can efficiently identify topics in large datasets and has been widely used in academic research. PLSA is a variation of LDA that can handle more complex structures in the data and has been shown to perform better than LDA in some cases. BERTopic is a more recent technique that utilizes pre-trained language models and has been shown to be highly effective in identifying latent topics in text data, especially in shorter and more informal texts. These techniques has the ability to extract meaningful topics from unstructured text data
- Model Evaluation: Different models implemented have different evaluation metrics as in the tables below. For Disaster Classification Result, ROC-AUC of SVM yielded the highest score, while for Disaster Type Classification, F1 score of Random Forest showed the highest score. BErtTopic appeared to be having the highest coherence score.

Disaster Classification				
Model	Precision	Recall	F1	<b><u>ROC-AUC</u></b>
Random Forest	0.730233	0.647156	0.686189	0.778145
<b>Support Vector Machine</b>	0.772018	0.825639	0.797951	<b>0.859081</b>
Feedforward Neural Network	0.702782	0.791426	0.744475	0.832107

Table 3. Disaster Classification Result

Disaster Type Classification				
Model	Accuracy	Precision	Recall	<b><u>F1</u></b>
<b>Random Forest</b>	0.852991	0.852916	0.852991	<b>0.850875</b>
Support Vector Machine	0.813373	0.860292	0.813373	0.824976
Feedforward Neural Network	0.822227	0.861992	0.822227	0.832449

Table 4. Disaster Type Classification Result

Topic modeling	
Model	<b><u>Coherence score</u></b>
LDA	0.315689
PLSA	0.283783
<b>BERTopic</b>	<b>0.623056</b>

Table 5. Topic Modeling Result

### 3.3.4 Potential Solutions

We strive to develop strategies to minimize the likelihood and impact of each risk. Here is one list of solutions how we faced these problems:

#### 1. Insufficient or low-quality data:

- Plan for additional data collection methods or sources
- Implement data preprocessing techniques to improve data quality
- Regularly monitor data collection progress

#### 2. Inaccurate or underperforming models:

- Use proven machine learning methods and algorithms
- Allocate more time for model development and testing

- Consider alternative approaches or techniques if needed

### 3. Integration challenges:

- Establish clear communication channels between developers and engineers
- Conduct regular integration tests to identify issues early
- Plan for additional time to address integration challenges

### 4. Technical difficulties or delays:

- Allocate contingency time in the project schedule
- Ensure team members have access to necessary tools and resources
- Provide training or support for team members as needed

### 5. Unexpected changes in project requirements or scope:

- Establish a change management process to handle scope changes
- Regularly communicate with stakeholders to clarify expectations
- Assess the impact of changes and adjust project plans accordingly

In addition, we also did risk monitoring and control in order to perfect our project.

1. Conduct regular risk assessments and update the risk management plan
2. Track the progress of risk mitigation strategies
3. Communicate risk updates and mitigation efforts with stakeholders
4. Adjust project plans as needed to address emerging risks

### **3.3.5 Tools Application**

Root cause analysis is a key tool used in this phase. The use of the Lean Six Sigma technique provides the necessary assistance to identify and deal with potential problems. It helps us effectively complete different kinds of problems and also keep phased problems and solutions that we may encounter until the next stage. During this process, we also utilized Trello to apply agile methodology, specifically Kanban approach in allocating our tasks.

## 3.4 Improve Phase

In the Improve phase, the focus is on identifying and prioritizing possible enhancements. This involves creating a plan for implementation, conducting a trial project, and evaluating the effectiveness of the implemented solution.

### 3.4.1 Solution Evaluation

Below are 2 essential areas for which alternative solution approaches were established, and the team examined the benefits and drawbacks of each.

**Data Preprocessing & Feature Selection Methods:** How we should preprocess Twitter texts and choose features before rolling into classification task.

Data Preprocessing & Feature Selection Methods	Pros	Cons
Stopword removal	Removing stopwords can help to reduce the amount of noise in the data, making it easier to identify important patterns and features.	It can sometimes result in the loss of important information, especially in cases where stopwords carry significant meaning or context.
Removing special characters and punctuations	Removing special characters and punctuation can simplify the text data and make it easier to process, especially for algorithms that are designed to work with alphabetic characters only	Special characters and punctuations sometimes carry important emphasis in the context and removing them can lead to the loss of important information.
Scaling numerical data using Standardization	It scales the data in a way that the mean of the data becomes 0 and the standard deviation becomes 1, making the model more robust to outliers.	It can be sensitive to the outliers in the data and may not perform well in certain types of data distributions, such as those with heavy tails.
Scaling numerical data using Min-Max Scaling	It scales the data to a fixed range (usually 0 to 1), making it easier to compare features with different units and ranges.	It can be sensitive to outliers and may not perform well when the distribution of the data is skewed or has a heavy tail.
PCA for Feature Selection	It can reduce the dimensionality of the data by identifying the most important features while still retaining most of the variation in the data.	It may not always be appropriate for non-linear relationships between features and can be computationally expensive for large datasets.
Chi Square Test	It can test association between	Data in Chi Square Test must be

	variables and identify differences between observed and expected values	numerical. Categories of 2 are not good to compare. The number of observations must be 20+ The test becomes invalid if any of the expected values are below 5
Exhaustive Feature Selection	It is the most robust feature selection method covered. It is a brute-force evaluation of each feature subset. It tries every possible combination of the variables and returns the best-performing subset.	Since it will try every combination, sometimes it will cost a lot of time to run the selection.
TF-IDF	Words that are more important for a given task will have higher TF-IDF scores, which can help to identify relevant features and reduce noise.	TF-IDF can be sensitive to the length of the document, as longer documents may have higher term frequencies and lower IDF scores, leading to a bias towards longer documents.
Word Embeddings	Word embeddings can capture the semantic relationships between words, allowing text classification models to better understand the meaning of the data. They can also be used on out-of-vocabulary words.	Words embeddings can be computationally expensive, especially for large datasets, since it involves training a neural network to learn the embeddings.
Tokenization	It helps to standardize the representation of text, making it easier to process and analyze.	Tokenization may lead to information loss, especially when dealing with phrases or idiomatic expressions.
Stemming and lemmatization	Stemming and Lemmatization can reduce the dimensionality of the data by reducing inflectional forms to their root form, which can improve the performance of text classification models.	Stemming can sometimes result in the loss of important information, especially in cases where inflectional forms carry significant meaning or context. While lemmatization can be computationally expensive, especially for larger datasets.

Table 6. Data Preprocessing &amp; Feature Selection Methods Analysis

**Model Selection & Evaluation:** Figuring out the most suitable approach for text classification and text clustering and how to evaluate their performances.

Model Selection & Evaluation	Pros	Cons
Support Vector Machines (SVM)	SVMs are effective for	SVMs can be sensitive to the



	high-dimensional data, and they work well with non-linearly separable classes by mapping the data to a higher dimensional space.	choice of kernel function and the regularization parameter, which can lead to overfitting or underfitting if not tuned properly. SVMs can also be computationally expensive for large datasets.
Mask R-CNN	Mask R-CNN is one of the most powerful deep learning models for object detection, allowing the model to identify and segment objects in an image, making it highly accurate.	The model is computationally expensive, requiring a lot of training data, and can be difficult to fine-tune and optimize.
Random Forest	Random Forest is a powerful ensemble method that can handle non-linear relationships and high-dimensional data. It is also less prone to overfitting than some other tree-based models.	It can be slow and computationally expensive to train, especially with a large number of trees and features.
Perceptron	It is a relatively simple and efficient algorithm that can be implemented quickly and trained on large amounts of data, particularly useful for binary classification tasks, such as spam filtering or sentiment analysis.	The perceptron algorithm is a relatively simple algorithm and may not achieve high accuracy compared to more complex methods, such as neural networks or ensemble methods.
Neural Networks	They are capable of learning complex relationships between the input features and the output classes, making them well-suited to text classification tasks that require sophisticated analysis of the text.	Neural networks typically require large amounts of data to train effectively, and may not perform well with small datasets.
Regression models	Regression models are used to predict a continuous target variable, which can be useful in a variety of applications and it can provide interpretable results that allow us to determine the effect of each predictor variable on the target variable.	Linear regression models assume a linear relationship between the predictor and target variables, which may not hold for more complex data. It can be sensitive to outliers in the data, which can result in a poor fit to the data.
KNN Clustering	KNN does not require any assumptions about the underlying data distribution,	KNN can be sensitive to the density of data in the feature space. Data points that are

	making it useful in a variety of applications. KNN can handle non-linear data, making it useful in applications where linear models are not appropriate.	located in low-density areas may be assigned to the wrong cluster.
RNN based models	RNN based models are designed to capture word dependencies and text structures, which can be helpful for understanding emergency tweets that may contain complex and nuanced language.	They can be computationally expensive to train and may suffer from the vanishing gradient problem, which can make it difficult to learn long-term dependencies.
CNN based models	CNN based models are trained to recognize patterns in text, which can be useful for identifying key phrases and information in emergency tweets.	They may not be able to capture long-term dependencies in text, which can be important for understanding the context and urgency of emergency tweets.
DL models with attention mechanism	Attention mechanism is effective at identifying correlated words in text, which can be helpful for understanding the relationships between different pieces of information in emergency tweets.	Attention models can be more complex than other DL models and may require more computation. Additionally, attention models may require more training data to achieve high accuracy compared to other models.
Precision score	It provides insight into how often positive predictions are correct, which can be important for decision-making.	It does not consider false negatives, meaning it may miss some true positive instances.
Recall score	Useful metric for evaluating classification models when the goal is to identify all positive instances	It does not consider false positives, meaning it may incorrectly label some negative instances as positive.
Accuracy score	It is easy to understand and interpret, and it is a popular choice for evaluating classification models.	It can be misleading when the class distribution is imbalanced, as a high accuracy score can be achieved simply by predicting the majority class.
F1 score	It is a good metric to use when the class distribution is imbalanced, as it gives equal weight to both positive and negative classes.	It can be difficult to interpret, especially when comparing models with different F1 scores.

Table 7. Model Selection &amp; Evaluation Analysis

### 3.4.2 Recommended Solution

Action Items	Allocated To	Delivery Date
Exploratory Data Analysis & Feature Selection	Ziyan Ping, Haoran Shu	02/27/2023
Model Selection & Evaluation	Duyen Nguyen, Praveen Allu	03/13/2023

Table 8. Data Analysis and Model Development

Action Items	Allocated To	Delivery Date
Model Performance Improvements	Duyen Nguyen, Praveen Allu	04/20/2023
Website Design and Implementation	Ziyan Ping, Haoran Shu	04/23/2023

Table 9. Model Evaluation and UI Design

### 3.4.3 Pilot Design

Our pilot phase has covered all aspects of the project, including data research, model construction and evaluation, and initial UI design. After the Voice of Consumer phase, we started developing the project based on customer input and expectations, which helped us stay focused throughout the various stages of the project. We had regular communication with our client or customer to keep them informed about the product's status and ensure that we were meeting their expectations.

To begin with, our machine learning classification models were constructed and tested using tweet data. We also built a topic model to determine the sentiment score of each topic, which helped us to assign a severity score to each filtered tweet based on its similarity to the topics and their severity scores. To find the most appropriate and accurate model, we tested with different models and compared their performance. Furthermore, we utilized Spacy to extract location information and also incorporated the time of the tweet to enhance the display of the end product.

Finally, we created a web-based UI interface prototype using Figma. This prototype included a map-based visualization that enabled the general public to view disasters, while also providing disaster management agents with informative alerts about potential disasters.

### 3.4.4 Work Breakdown Structure

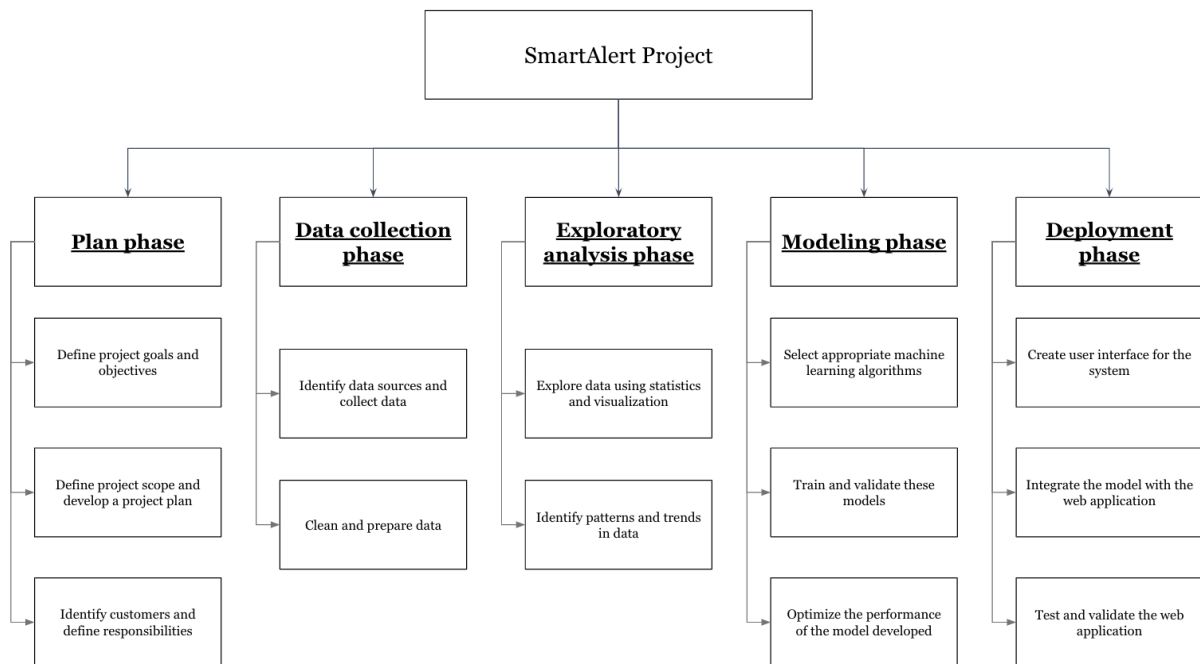


Figure 1. Work Breakdown Structure

The Work Breakdown Structure is designed to efficiently allocate tasks and resources to ensure successful completion of the project. The primary goal of the project is to develop a web application that leverages social media data, specifically from Twitter, to provide timely and accurate alerts during natural disasters. The WBS breaks down the project into manageable components, enabling the team to focus on each aspect of the development process.

Our WBS consists of five main components: Plan Phase, Data Collection Phase, Exploratory Analysis Phase, Modeling Phase and Deployment Phase.

- **Plan Phase:** tasks such as planning, resource allocation, and progress monitoring are handled. This ensures the smooth functioning of the project and proper communication among team members.
- **Data Collection Phase:** appropriate data sources were identified and data was collected. The collected data was then cleaned and prepared to ensure its suitability for the subsequent phases of the project.
- **Exploratory Analysis Phase:** we explored the data using statistical analysis and visualization techniques. This allowed us to identify patterns and trends within the data, which helped to inform our decision-making process throughout the project. By carefully

---

analyzing the data, we were able to gain valuable insights that informed the development of our machine learning models and ultimately led to the successful completion of the project.

- **Modeling Phase:** appropriate machine learning algorithms were selected and used to train and validate the models. The performance of the developed models was optimized through various techniques, including hyperparameter tuning and feature engineering. The final models were selected based on their accuracy and efficiency in processing real-time tweets related to natural disasters.
- **Deployment Phase:** we created a user interface for the system, integrated the model with the web application, and tested and validated the web application. The user interface was designed to be user-friendly and intuitive, allowing authorized users to easily access the system and perform necessary tasks. The model was integrated seamlessly with the web application, ensuring that it functioned smoothly and accurately. We also conducted thorough testing and validation to ensure that the web application met all requirements and functioned as expected.

## 3.5 Control Phase

The control phase is a data-driven approach used for process improvement and reducing variability in a system. After defining the problem, measuring the current performance, analyzing the root causes, and implementing improvements, the control phase aims to ensure that the improvements are sustained over time, preventing the process from reverting to its previous state.

To implement and sustain the solutions that were developed during the Improve phase, we monitor and track the effectiveness of the solutions and make adjustments as necessary. We also develop a communication plan to keep stakeholders informed of any updates or changes to the solution.

### 3.5.1 Control Solutions Considered

We explored the following two methods to implement guardrails that would further improve the process:

1. **User Experience (UX) Optimization:** Focus on minimizing manual input from end-users, as each interaction point presents an opportunity for errors to be introduced. Design the app's user interface to be intuitive and easy to use, incorporating features like error prevention and form validation.
2. **Content Management System (CMS) Integration:** On the back-end, use a CMS to ensure that data is always in sync, and any updates are reflected on the front-end as soon as they are made on the back-end. Provide fallback mechanisms in case of issues to maintain a seamless experience for users.

### 3.5.2 Control Solution Implemented

In order to maintain high performance of the deliverables, the following measures of control have been implemented:

1. **User Experience (UX) Optimization:** We have an "Add Alert" button, exclusively available to authorized emergency management agents, which enables them to input emergency information if it is reported through other media sources. This alert will be displayed on the "Review Alerts" section, along with other alerts sourced from Twitter, awaiting verification from agents. Once the emergency has been verified, it will be shown on the map.

2. **Content Management System (CMS) Integration:** We utilized firebase as one medium for storing location information. In this way, it will have one immediate update when there is one new emergency added.

## 3.6 Result and System Implementation

By utilizing the DMAIC methodology, the team was able to streamline processes, resulting in cost savings and quicker project completion. The system implemented and its corresponding results are summarized below.

### 3.6.1 Machine Learning Approaches

#### 1. Disaster-related tweet classification model:

- We extracted features from our tweet texts using Word2Vec, specifically, the word2vec-google-news-300 pre-trained model.
- Since our dataset is imbalanced, with class 0 (non-emergency-related tweets) being 60% more than class 1 (emergency-related tweets), we used SMOTE to create synthetic data to balance out the dataset.
- As shown in table ..., we ran 3 classification models, random forest, support vector machine, and feedforward neural network to see which one would yield the highest AUC-ROC score. SVM was chosen as our final classification model.

#### 2. Disaster type classification model

- We used the same dataset used above but instead of using disaster\_label as our target feature, we used disaster\_type to train our classifier models.
- We ran four classification models, random forest, SVM, feed forward network and logistic regression to see which one would yield the highest metric score.
- F1 score was chosen as an evaluation metric to evaluate these models, we found that the Random forest model yielded the best results with the highest score. As shown in table ..., we chose RF as our final disaster type classification model.

#### 3. BertTopic model

- The model uses a pre-trained DistilBERT model to extract features from text data.
- The UMAP algorithm is used for dimensionality reduction, while HDBSCAN is used for clustering. We defined a seed\_topic\_list consisting of predefined topics



related to disasters. The BERTopic model is then trained on the disaster-related tweets, and the output consists of a list of topics and their corresponding probabilities.

- We followed the similar methodology to create topics models based on LDA and PLSA.
- We used coherence score as a metric to evaluate the topic models. This metric measures how interpretable the topics are. It is based on the idea that a good topic model should produce topics that are coherent and meaningful. Coherence is calculated by measuring the degree of semantic similarity between the top words in a topic.
- We used a sentiment analysis tool such as SentiWordNet to calculate sentiment scores for each word in the extracted top N words. We calculate the sentiment score for each topic by taking the average of the sentiment scores of the words in the topic. This gives us an overall sentiment score for each topic. In the later step, we assign a severity score to each topic based on the sentiment score. Finally, we calculate a severity score for each filtered tweet based on the similarity of the tweet text to the topics and the severity scores of the topics.

#### 4. Spacy, Geopy model

- We load a pre-trained spaCy model to identify locations in the tweets. The geopy library is then used to geocode the location data, which is used to generate latitude and longitude coordinates for each tweet.

### 3.6.2 System Implementation

- HDPZ's goal is to improve response times during natural disasters by developing a system that can quickly notify disaster management agents and the general public with accurate information. This will be accomplished by using machine learning algorithms to identify and categorize relevant Tweets related to natural disasters by type, severity level, and location. Once verified, this information will be disseminated to the appropriate parties, helping to minimize damage and save lives.
- This project is divided into 5 steps in order to build an efficient natural disaster alert system.

- Firstly, we used Word2Vec for feature extraction from tweet texts and then tested three classification models - random forest, support vector machine, and feedforward neural network, utilizing AUC-ROC as the evaluation metric. With the lowest AUC-ROC yielded, we chose SVM as our final classification model with the objective of identifying if the incoming tweet pertains to a disaster or not. Once identified, disaster-related tweets will be processed to the next step.
- Secondly, disaster-related tweets are processed to classify the disaster type using a random forest model with F1 score as the evaluation metric.
- Then, topic modeling is performed using BERTopic model and coherence score as the evaluation metric. Sentiment analysis is conducted to calculate the sentiment score for each topic, which is then used to assign a severity score to each topic. Finally, a severity score is calculated for each filtered tweet based on the similarity of the tweet text to the topics and the severity scores of the topics.
- After performing the above steps, location data is extracted from tweets using Spacy and Geopy model to generate latitude and longitude coordinates for each tweet.
- Lastly, we developed a web application that serves as the primary interface for the SmartAlert system, making it easily accessible to disaster management agents, government agencies, and the general public.
  - The web application is designed with a user-friendly interface and includes features such as a map for displaying real-time natural disaster events, an alerts management system for disaster management agents, and information tabs to educate users about the project and its objectives.
  - The implementation of the SmartAlert system starts with deploying the web application on a secure and scalable cloud infrastructure of Fire Base, ensuring that it is capable of handling a large volume of data. The system will be regularly updated with the results yielded from our AI model that processes tweets extracted from Apify scraper in order to maintain its accuracy and effectiveness in identifying, categorizing, and disseminating relevant natural disaster information.

### 3.6.3 Prototype and Demonstration

The prototype of the SmartAlert system is a fully functional web application, demonstrating the core features and capabilities of our natural disaster alert solution. In this section, we will provide a brief overview of the key components of the prototype, which will be accompanied by screenshots to help visualize the system's design and functionality.

- Disaster Map:
  - The Disaster Map is a central feature of the SmartAlert system, providing users with a real-time, interactive display of ongoing natural disaster events. The map is designed with user-friendly icons and color-coded markers to easily differentiate between various disaster types, severity levels, and locations. Users can click on individual markers to view detailed information about each event, such as its type, severity, and the source of the alert.

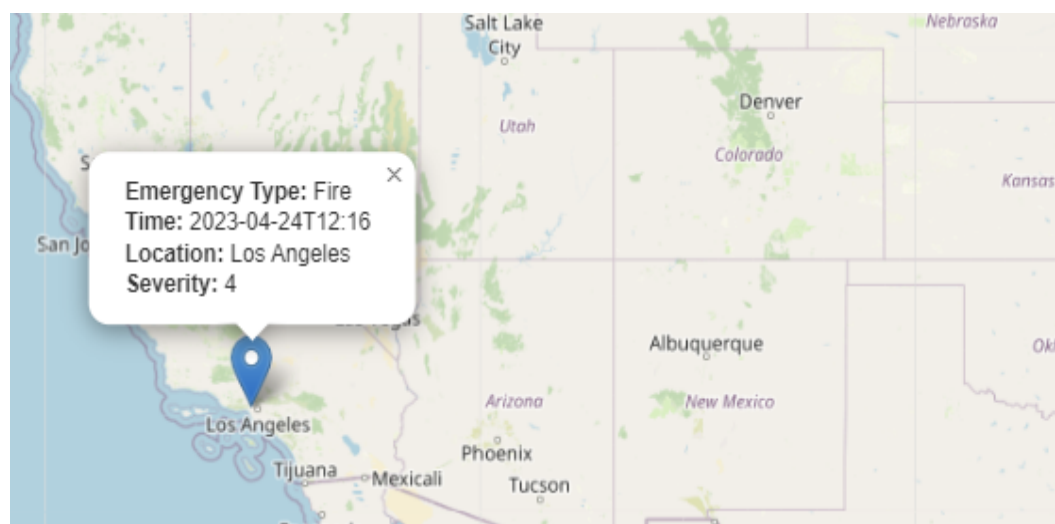
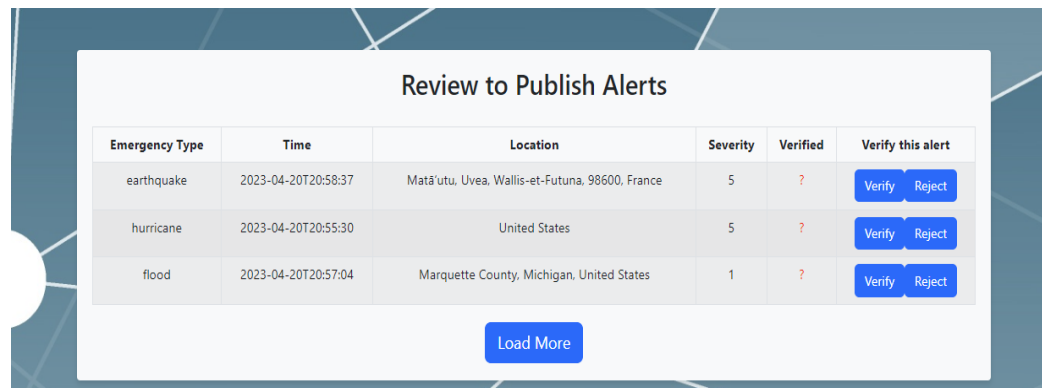


Figure 2. Disaster Map

- Alert Review:
  - This component is designed for disaster management agents, who can access a dedicated interface for reviewing potential natural disaster events detected by our AI model. The Alert Review tab displays location, time, and severity prediction for each event, allowing agents to verify the accuracy of the information before publishing it on the Disaster Map. This ensures that only reliable and up-to-date information is disseminated to the public.

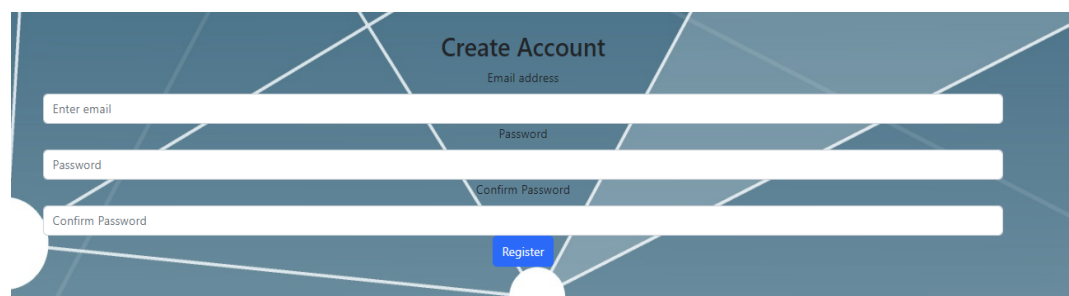


Emergency Type	Time	Location	Severity	Verified	Verify this alert
earthquake	2023-04-20T20:58:37	Matā'utu, Uvea, Wallis-et-Futuna, 98600, France	5	?	<button>Verify</button> <button>Reject</button>
hurricane	2023-04-20T20:55:30	United States	5	?	<button>Verify</button> <button>Reject</button>
flood	2023-04-20T20:57:04	Marquette County, Michigan, United States	1	?	<button>Verify</button> <button>Reject</button>

Load More

Figure 3. Alert Review

- Users Management:
  - The Users Management feature allows administrators to efficiently manage and oversee user accounts, including those of disaster management agents and general users. Administrators can add, edit, or delete user accounts, assign roles and permissions, and monitor user activity to ensure the system's integrity and security.



Create Account

Email address

Enter email

Password

Password

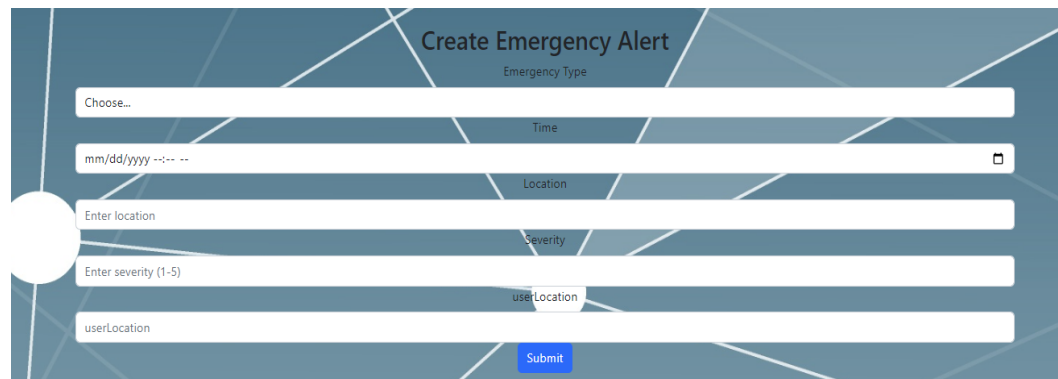
Confirm Password

Confirm Password

Register

Figure 4. Users Management

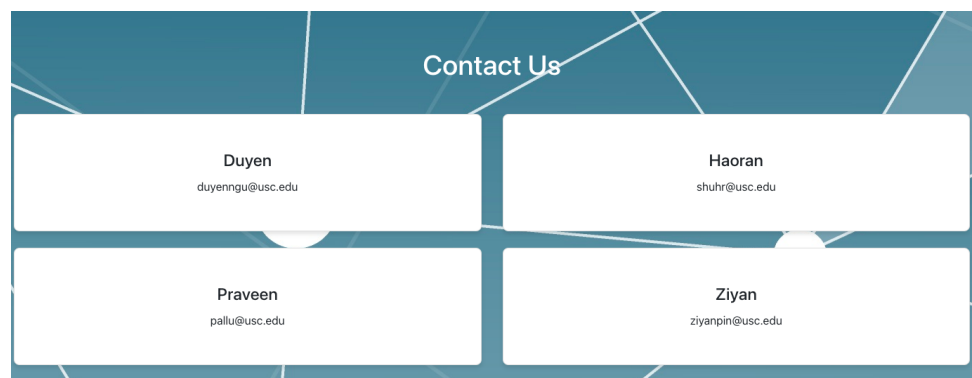
- Alert Creation Form:
  - The Alert Creation Form is a specialized tool for disaster management agents to manually create disaster alerts sourced from other media. Agents can input critical information about the event, such as its type, location, severity, and description. Once submitted, these alerts will appear directly on the Disaster Map, providing users with the most current and accurate information available.



The screenshot shows a web form titled "Create Emergency Alert". The form has a blue header with the title. Below the header, there are five input fields: "Emergency Type" (a dropdown menu with "Choose..." as the placeholder), "Time" (a date and time picker showing "mm/dd/yyyy --:-- --"), "Location" (a text input field with "Enter location" as the placeholder), "Severity" (a text input field with "Enter severity (1-5)" as the placeholder), and "userLocation" (a text input field with "userLocation" as the placeholder). A blue "Submit" button is located at the bottom right of the form.

Figure 5. Alert Creation Form page

- Contact Us:
  - Contact Us provides our email addresses for further concerns and questions regarding our project.



The screenshot shows a web page titled "Contact Us". The page has a blue header with the title. Below the header, there are four white rectangular boxes arranged in a 2x2 grid. Each box contains the name of a team member and their email address: Duyen (duyengu@usc.edu), Haoran (shuhr@usc.edu), Praveen (pallu@usc.edu), and Ziyan (ziyanpin@usc.edu).

Figure 6. Contact Us page

- About:
  - About page displays the members of HDPZ, our project summary and how to use the web application.

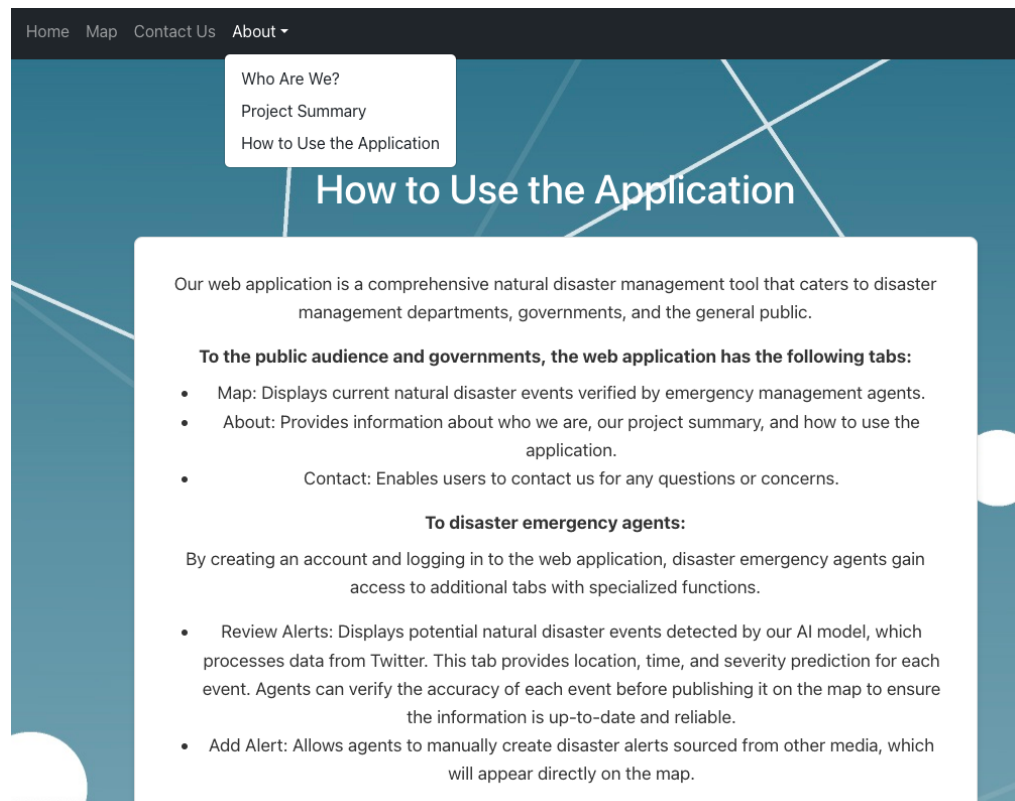


Figure 7. About page

## References

1. Mohamed Bakillah, Ren-Yu Li & Steve H.L. Liang (2015) "Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan", International Journal of Geographical Information Science, 29:2, 258-279

<https://www.tandfonline.com/doi/abs/10.1080/13658816.2014.964247>

2. Klein, B., Laiseca, X., Casado-Mansilla, D., López-de-Ipiña, D., Nespral, A.P. (2012).

Detection and Extracting of Emergency Knowledge from Twitter Streams. In: Bravo, J., López-de-Ipiña, D., Moya, F. (eds) Ubiquitous Computing and Ambient Intelligence. UCAmI 2012. Lecture Notes in Computer Science, vol 7656. Springer, Berlin, Heidelberg.

[https://doi.org/10.1007/978-3-642-35377-2\\_64](https://doi.org/10.1007/978-3-642-35377-2_64)

3. Armin Mahmoodi, Milad Jasemi Zergani, Leila Hashemi, Richard Millar. "Analysis of optimized response time in a new disaster management model by applying metaheuristic and exact

methods”, Smart and Resilient Transportation. (n.d.).

<https://www.emerald.com/insight/content/doi/10.1108/SRT-01-2021-0002/full/html>

4. AminiMotlagh, M., Shahhoseini, H. & Fatehi, N. A reliable sentiment analysis for classification of tweets in social networks. Soc. Netw. Anal. Min. 13, 7 (2023).

<https://doi.org/10.1007/s13278-022-00998-2>

## Appendix

Suppliers	Inputs	Process	Outputs	Customers
<ul style="list-style-type: none"> <li>Data Sources (Twitter API, Kaggle, Crisisnlp.qcri.org)</li> <li>Safety departments (police department, fire department, EMS,...)</li> <li>Collaboration tools (Github, Google Drive, Zoom)</li> <li>USC Library for related works</li> </ul>	<ul style="list-style-type: none"> <li>Tweets</li> <li>List of safety departments</li> <li>Contact information of local residents</li> <li>ML technical skills</li> </ul>	<ul style="list-style-type: none"> <li>Data Acquisition</li> <li>Data Exploration</li> <li>Feature Extraction</li> <li>Model Classification</li> <li>Model Clustering</li> <li>Model Prediction</li> <li>Web-app design</li> <li>Linking web-app with prediction model</li> <li>Web-app deployment</li> </ul>	<ul style="list-style-type: none"> <li>Emergency event detection</li> <li>Emergency severity level prediction</li> <li>Appropriate department assignment</li> <li>Web application</li> <li>Documentation</li> </ul>	<ul style="list-style-type: none"> <li>Local residents</li> <li>Safety departments (police department, fire department, EMS,...)</li> <li>Government</li> </ul>

Figure 8. SIPOC map

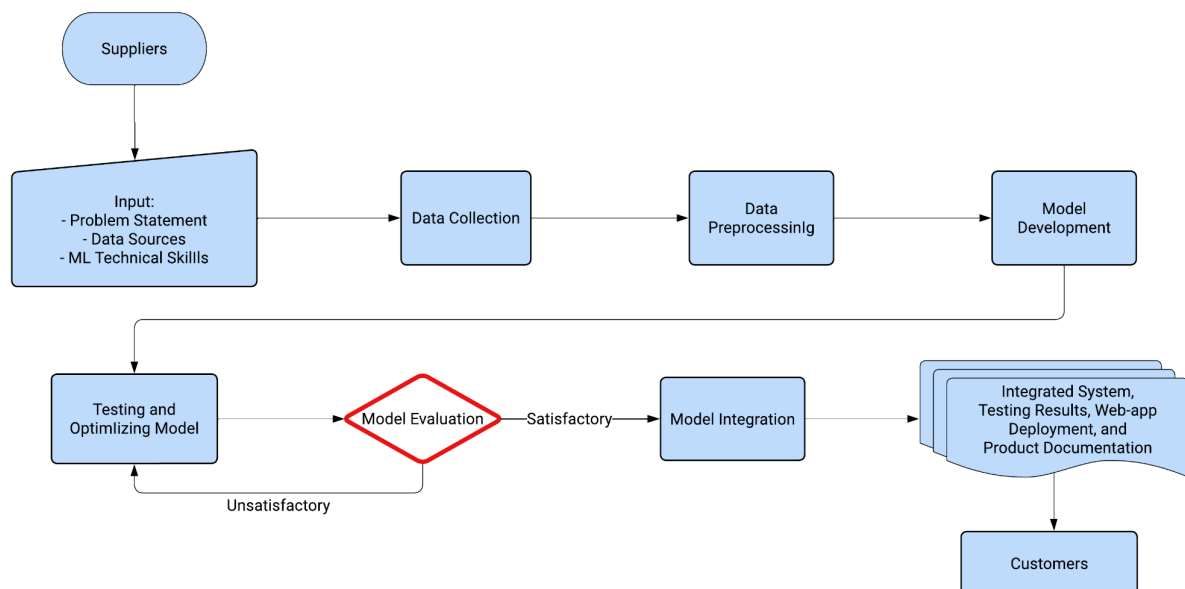


Figure 9. Common Process Map

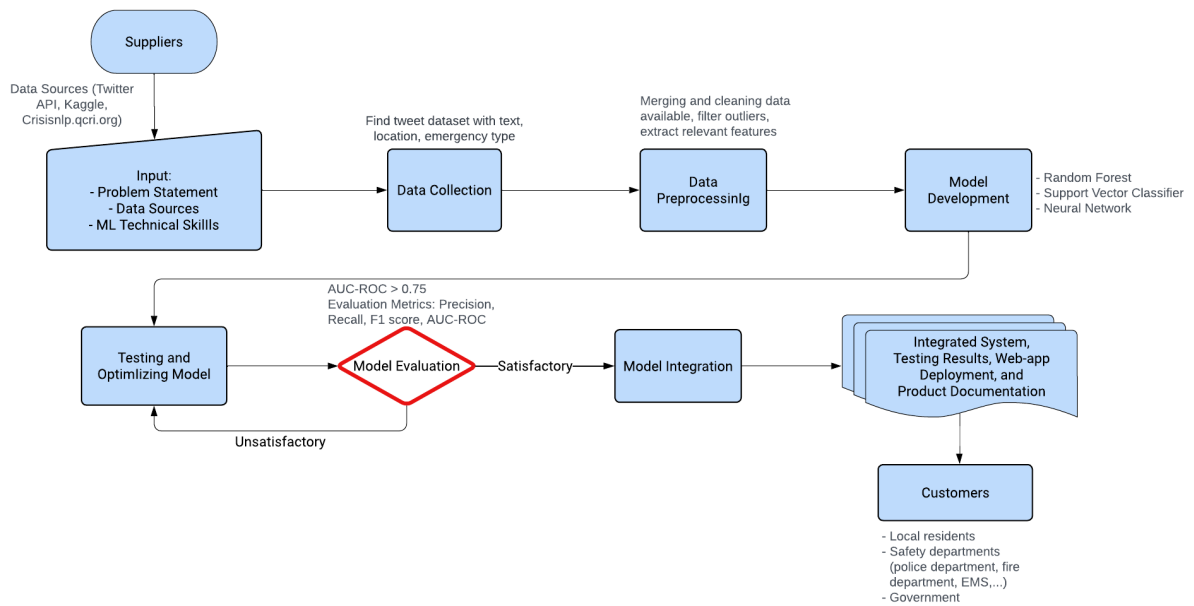


Figure 10. Detailed Process Map

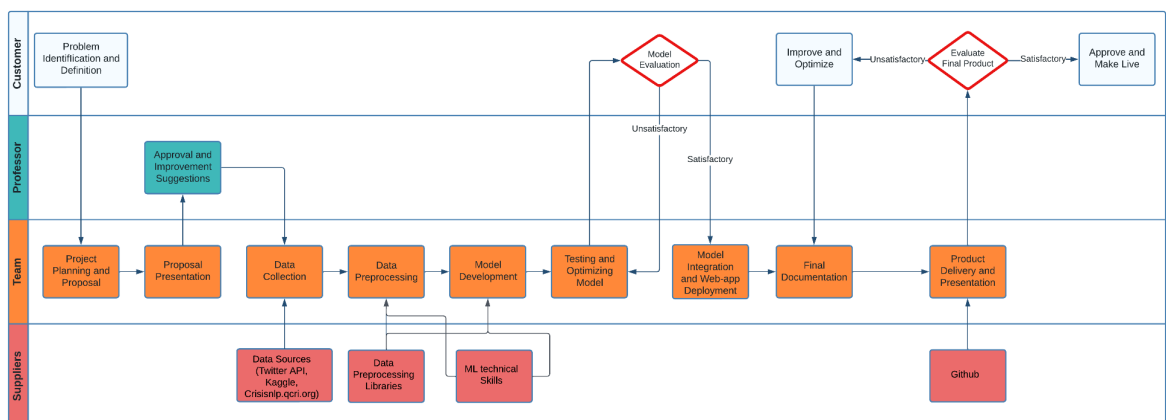


Figure 11. Functional Process Map



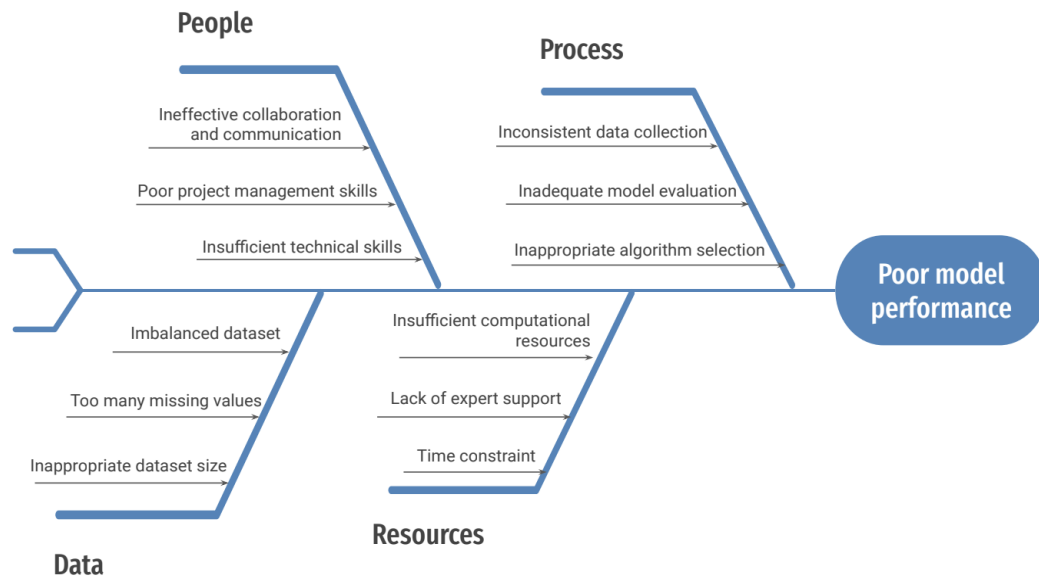


Figure 12. Fishbone diagram

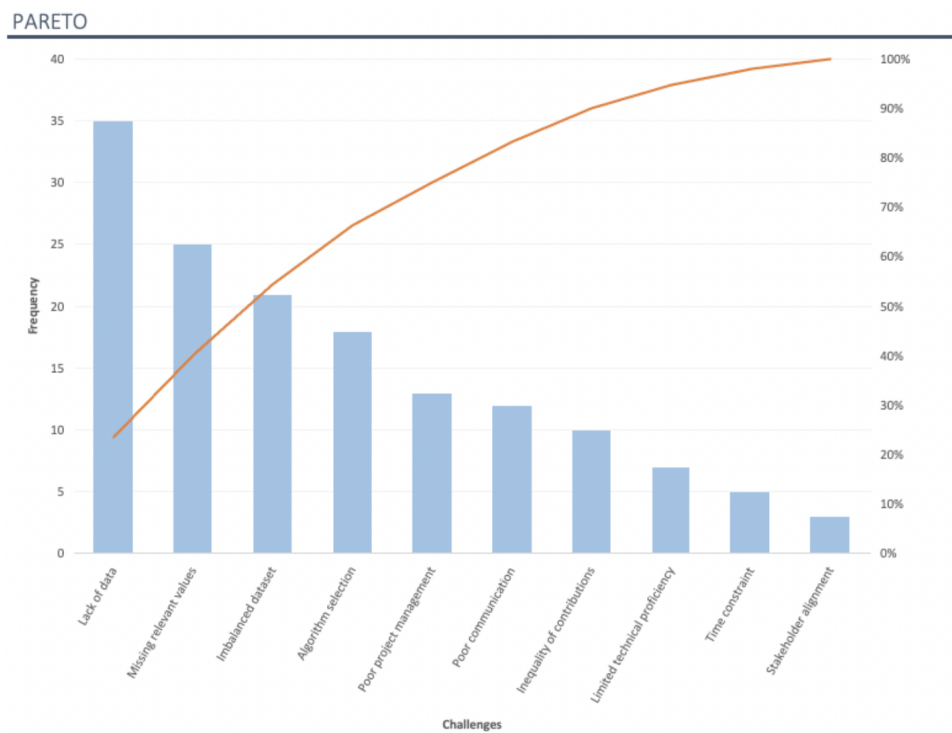


Figure 13. Pareto's chart