Project Proposal

AKADEMI | PHASE 5

PREPARED BY:
- AMEE H. JEUDY
- WOODNALIE S. JOSPEH

OCTOBER 2025

Table of contents

Executive Summary

Our project focuses on developing a Natural Language Processing (NLP) model that can automatically classify disaster-related text messages into relevant categories. During emergencies, organizations receive large volumes of messages from the public that may contain urgent requests, reports of damage, or general information. Manually sorting through these messages can delay critical response actions. Our goal is to create a model that supports faster and more accurate message categorization, allowing disaster response teams to prioritize resources more effectively.

We are working with a dataset of disaster messages provided by Figure Eight, which includes both messages and their corresponding categories. The data represents real-world communication during various disaster events, making it ideal for building a model that can generalize across multiple contexts.

The project applies standard NLP preprocessing techniques such as tokenization, lemmatization, and stop word removal, followed by feature extraction using TF-IDF. We plan to train and evaluate several supervised learning models, starting with a logistic regression baseline and progressing to more complex algorithms such as random forest or gradient boosting, to identify the best-performing classifier.

Success will be measured using multi-label classification metrics such as F1-score, precision, recall, and accuracy. Our aim is to build a model that achieves strong recall for critical categories, minimizing the risk of missing important messages that indicate human need or danger.

Although we do not plan to deploy a web application, the final deliverable will be a well-documented Jupyter notebook that demonstrates the full end-to-end workflow: data preparation, model training, evaluation, and interpretation of results. This approach ensures transparency and reproducibility, enabling others to build upon our work or adapt the model for use within an operational disaster response system.

Through this project, we hope to contribute a small but meaningful step toward leveraging machine learning for humanitarian purposes, where time-sensitive information processing can directly impact lives and communities in crisis.

Business understanding

WHAT ?

The project aims to develop a natural language processing (NLP) model capable of classifying messages from humanitarian communication channels into two categories: urgent requests for aid and informational content. In disaster contexts such as the 2010 Haiti earthquake, rapid identification of aid requests can significantly improve response coordination and resource allocation.

WHY ?

In humanitarian crises, communication overload can delay critical interventions. Filtering incoming messages based on urgency helps prioritize resources and save lives. This topic matters because efficient crisis communication is central to disaster management, and the proposed system supports this mission. The project is personally relevant because it focuses on Haiti where disaster communication infrastructure faces chronic challenges.

FOR WHO?

This work sits at the intersection of crisis informatics, humanitarian AI, and ethical NLP. It applies directly to international NGOs (Red Cross, UN OCHA), emergency management agencies, and disaster tech platforms like WhatsApp crisis-response systems.

STAKEHOLDERS

- The primary audience is humanitarian field coordinators and triage teams who must prioritize aid under extreme time pressure.
- Secondary audiences include disaster-response policymakers, and platform developers integrating NLP into humanitarian workflows.

IMPACT

If implemented, the model could automatically flag messages that indicate immediate human needs such as requests for food, water, shelter, or medical aid. This would allow responders to act faster, allocate resources more effectively, and potentially reduce loss of life. The approach could also be integrated into existing humanitarian communication platforms to improve real-time response management.

PRIOR WORK

We build on the Figure Eight Disaster Response Messages dataset, used in Kaggle competitions and academic papers. Prior research in disaster response NLP (e.g., Imran et al., 2014; Alam et al., 2021) demonstrates the value of supervised models in classifying humanitarian text. This project extends those ideas by applying them specifically to Haitian disaster communications, enhancing interpretability and practical deployment potential.

Data understanding

The data for this project comes from the Figure Eight Disaster Response Messages dataset, which contains thousands of messages sent during various global disasters. For this project, we are going to focus on the subset related to the 2010 Haiti earthquake.

The dataset is publicly available on Kaggle and consists of two main CSV files:

- messages.csv (containing the text and metadata) and
- categories.csv (containing the message labels).

The two datasets were merged on the id field to create a single dataset for modeling.

Key features include:

- message: the text content of the communication (main input).
- genre: the channel type, such as direct, news, or social.
- category columns: binary indicators representing various message types (food, water, medical_help, aid_related).
- For this project, only the binary classification between urgent request and informational message was used as the target variable.

Other researchers have used this dataset for broader classification tasks involving multiple labels. This project builds on that foundation but narrows the focus to detecting urgency, which makes it directly applicable in operational settings where decision speed is essential.

The data is already structured and accessible, requiring no additional collection effort. It provides an adequate volume for training and testing reliable models.

Data preparation

DATA SOURCE AND COLLECTION PLAN

The dataset is stored in CSV format and processed in Python using pandas. Each record includes text and categorical fields. The message column is of type string, the genre is a categorical variable, and the target variable (request) is binary (0 or 1). After merging and cleaning, the dataset contained roughly 26,000 rows, with a near-balanced class distribution.

PREPROCESSING STRATEGY AND CHALLENGES

The preprocessing pipeline included several key steps:

1. Converting text to lowercase
2. Removing punctuation, special characters, and extra spaces
3. Tokenizing sentences into words
4. Removing stopwords in both English and Haitian Creole
5. Lemmatizing words using NLTK's WordNetLemmatizer
6. Transforming the cleaned text into TF-IDF features using unigrams, bigrams, and trigrams
7. Applying a stratified train-test split (80/20)

We anticipate several challenges during data preparation. The dataset contains messages written informally, often with abbreviations, mixed languages, and spelling variations. Managing these inconsistencies will be important to ensure that the model learns the right linguistic patterns. Another challenge will be handling the sparsity of high-dimensional TF-IDF matrices without losing interpretability.

We will also use exploratory visualizations such as word clouds, bar charts, and frequency plots to better understand message patterns, word distributions, and label proportions. These visual analyses will help verify that the data behaves as expected before moving to the modeling phase.

Modeling

We will frame this as a supervised binary classification problem, where each message will be classified as either “urgent” or “informational.”

For the baseline model, we will start with Logistic Regression because it is simple, efficient, and interpretable. Logistic Regression works well with sparse TF-IDF features and provides clear insight into which words most strongly influence predictions.

Once we establish the baseline, we will experiment with additional algorithms such as Multinomial Naive Bayes, Linear SVM, and possibly Random Forest for comparison. We will evaluate these models to determine which best balances performance, interpretability, and computational cost.

We plan to perform hyperparameter tuning using grid search or randomized search to optimize model parameters such as regularization strength and n-gram range.
Our modeling phase will also include feature analysis, where we identify the most informative words that indicate urgent requests. This step will make the model’s behavior more transparent and useful for humanitarian interpretation.

The model pipeline will include TF-IDF vectorization, feature scaling where appropriate, and stratified splitting to maintain balanced class representation during training and testing.

Evaluation

We will evaluate our models using a combination of performance metrics: accuracy, precision, recall, F1-score, and ROC-AUC. Since the cost of missing an urgent message is higher than the cost of flagging an informational one, recall will be our primary metric.

Our minimum viable product (MVP) will be a functioning binary classifier that achieves strong recall and balanced F1-score on a held-out test set. The MVP goal will be to demonstrate that NLP-based text analysis can reliably identify urgency in real disaster messages.

Once the MVP is achieved, we will work on stretch improvements, such as:

- Optimizing feature extraction and model parameters for higher performance.
- Analyzing misclassified examples to understand the model's limitations.
- Creating interpretability visualizations to show which words or phrases most strongly drive predictions.
- Conducting cross-validation to confirm model stability across multiple subsets.

By the end of the evaluation stage, we expect to have a model that performs well enough to be trusted for informational triage, accompanied by clear explanations and visual summaries of its behavior.

Deployment

We plan to report our final results through a structured Jupyter notebook and a clear project presentation. The notebook will contain our full workflow from data loading and cleaning to model evaluation, along with visualizations and concise explanations that guide the reader through our reasoning process.

While we do not intend to develop a web application, our goal is to make the analysis reproducible and easily interpretable. The notebook will serve as a transparent and self-contained report that demonstrates how the model can be applied to real-world disaster response data.

The final functionality focuses on taking incoming disaster-related text messages, preprocessing them using the same cleaning pipeline, and classifying them into multiple categories (such as “request,” “medical help,” “infrastructure,” etc.). This setup shows how the model could be integrated into a disaster response pipeline, even if the deployment interface is not built.

Our presentation will include key findings, model performance summaries, confusion matrices, and example predictions to help stakeholders understand how this model could assist in prioritizing disaster response efforts in a real-world setting.

Tools and methodology

We will use Python as our main programming language and work in a Jupyter Notebook environment. The data will be stored locally as CSV files.

Key tools and libraries include:

- pandas, numpy for data manipulation and analysis
- nltk, re, string for text cleaning and preprocessing
- scikit-learn for TF-IDF vectorization, model training, and evaluation
- matplotlib, seaborn, and wordcloud for visualization and communication of results

The workflow will follow a structured data science process:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation

We will complete all analysis locally, ensuring that our workflow remains reproducible and efficient. The final deliverable will include a clear and well-documented notebook with all steps and results explained, supported by visual insights and evaluation summaries.

Contact:

ameehashleyjeudy@gmail.com
linkedin.com/in/amee-hashley-jeudy-460449325

woodnaliesjoseph@gmail.com
linkedin.com/in/woodnalie-s-joseph-aa0011175

Octobre 2025