Disaster Tweet Analyzer

Introduction:

The aim of this work is to employ data analytic techniques to build a classification model that can accurately identify tweets referring to real disasters[1]. The **Disaster Tweet Analyzer** project leverages Natural Language Processing (NLP) techniques to classify tweets as disaster-related or non-disaster-related, providing crucial support during emergencies. In times of crisis, social media platforms like Twitter become critical for real-time information exchange, but the sheer volume of data makes it difficult to extract actionable insights. This project aims to automate the identification of disaster-related tweets, helping emergency responders and organisations quickly filter out essential information. By analysing tweet content, keywords, and locations, the system enhances crisis communication and improves the efficiency of disaster response efforts.

Dataset and Methodology:

The dataset for the Disaster Tweet Analyzer project is sourced from the Kaggle[2]. The dataset includes a file named tweets.csv, which has a size of 1.62MB and contains 11,370 rows.

Features:

- id: A unique identifier for each tweet.
- text: The content of the tweet.
- location: The location from which the tweet was sent (this field may be blank).
- keyword: A keyword associated with the tweet (this field may be blank).
- target: Denotes whether a tweet is about a real disaster (1) or not (0).

The methodology for the Disaster Tweet Analyzer project includes the following key steps up to Week 2:

1. **Data Collection**: The dataset is sourced from Kaggle and consists of tweets related to disasters, titled "Disaster Tweets."
2. **Data Preprocessing**: This step involves cleaning the data by handling missing values, removing URLs, managing emojis and tags, eliminating unnecessary characters, removing stopwords, and applying stemming.
3. **Feature Engineering**: Vectorizing tweet text using TF-IDF or Countvectorizer.

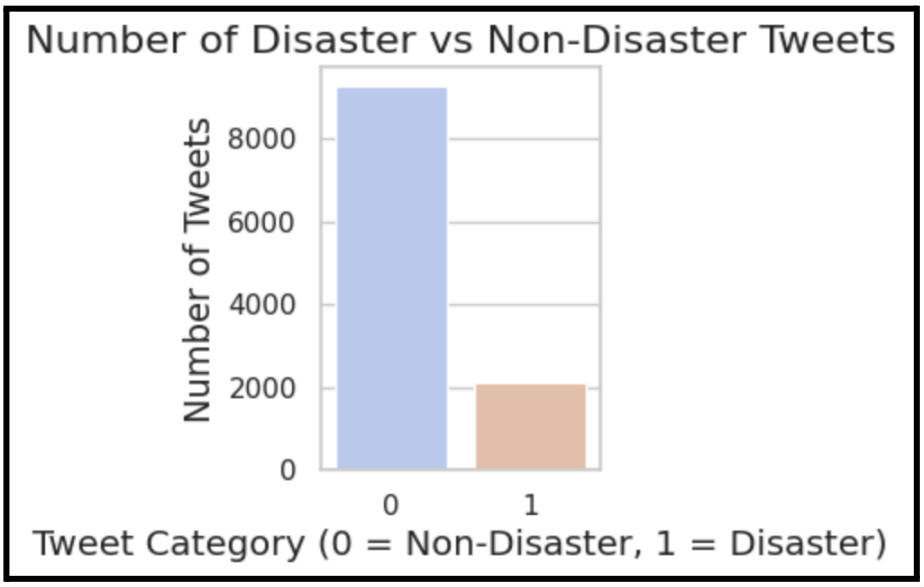
This is the methodology used as of now.

Results:

The Disaster Tweet Analyzer project has yielded the following key results through the initial analysis phases:

Analysing the data:

- Checking for the null values.
- Visualising relationships
- Visualising most frequent words using word cloud



Preprocessing the data:

- Removing URLs:
  - Before removing URL:

    #ThankfulTuesday Isaiah 43:2 When you pass through the waters, I will be with you; and when you pass through the r… https://t.co/jEp3L1oU36

  - After removing URL:

    #ThankfulTuesday Isaiah 43:2 When you pass through the waters, I will be with you; and when you pass through the r…

- Removing emojis:
  - Before Handling Emojis:

    pre-order untuk Map of the Soul: 7 oleh ARMY China 🇨🇳 telah mencapai 230.192 copy 😮gileee gileee #BestFanArmy #BTSARMY #…

  - After Handling Emojis:

    pre-order untuk Map of the Soul: 7 oleh ARMY China :flag: China: telah mencapai 230.192 copy :face with open mouth:gileee gileee #BestFanArmy #BTSARMY #…

- Removing Useless characters:
  - Before Removing Useless Character:

    Rengoku sets my heart ablaze:pensive face::red heart::fire: P.s. I missed this style of coloring I do so here it is c: #鬼滅の刃

  - After Removing Useless Character:

    Rengoku sets heart ablazepensive facered heartfire missed this style coloring hereAnalysing the data:

- Removing stopwords and stemming:
  - Before Removing Stopwords:

    Rengoku sets heart ablazepensive facered heartfire missed this style coloring here

  - After Removing Stop Words and Stemming:
    rengoku set heart ablazepens facer heartfir miss style color

Conclusion:

The Disaster Tweet Analyzer project utilises Natural Language Processing techniques to identify tweets related to disasters. By processing and analysing tweet content, the model aims to improve the efficiency of disaster response efforts. Predicting tweets as a disaster or not can help in real time rescue operations during a disaster, for fundraising to provide relief to affected people during a disaster and can also help in reporting of such events[3].

Future objectives for next two weeks:

For the next two weeks, the objectives could include:

- Refining the feature engineering process using more advanced techniques (e.g., word embeddings).
- Training machine learning models for classification.
- Evaluating model performance through metrics like accuracy, precision, and recall.
- Iterating on the preprocessing steps to improve model results.
- Implementing a more robust disaster classification system to aid real-time responses.

References:

[1] https://actascientific.com/ASCS/pdf/ASCS-05-0420.pdf

[2]https://www.kaggle.com/datasets/vstepanenko/disaster-tweets?resource=download&select=tweets.csv

[3] https://www.irejournals.com/formatedpaper/1704031.pdf