**Disaster Tweet Analyzer: Natural Language Processing for Crisis Communication**

In the age of social media, real-time data from platforms like Twitter has proven invaluable for understanding and responding to crises. The **Disaster Tweet Analyzer** project leverages Natural Language Processing (NLP) techniques to analyse tweets in real-time, filtering those that are relevant to disasters. By focusing on tweet classification, this tool aims to support disaster response teams by quickly identifying actionable information.

The project involves gathering, cleaning, and processing tweet data, followed by the application of machine learning algorithms to distinguish between disaster-related and non-disaster tweets. The insights derived from this system could help responders prioritize resources and provide timely assistance during emergencies.

**1. Introduction**

This project focuses on building a tweet classification model to support disaster response using machine learning techniques. The key goal is to differentiate between tweets that are related to disaster events and those that are not. This classification will help in the early identification of relevant tweets, providing timely information for disaster management.

The report summarizes the work done in Weeks 1 and 2 of the project, including dataset exploration, preprocessing steps, and initial feature extraction efforts.

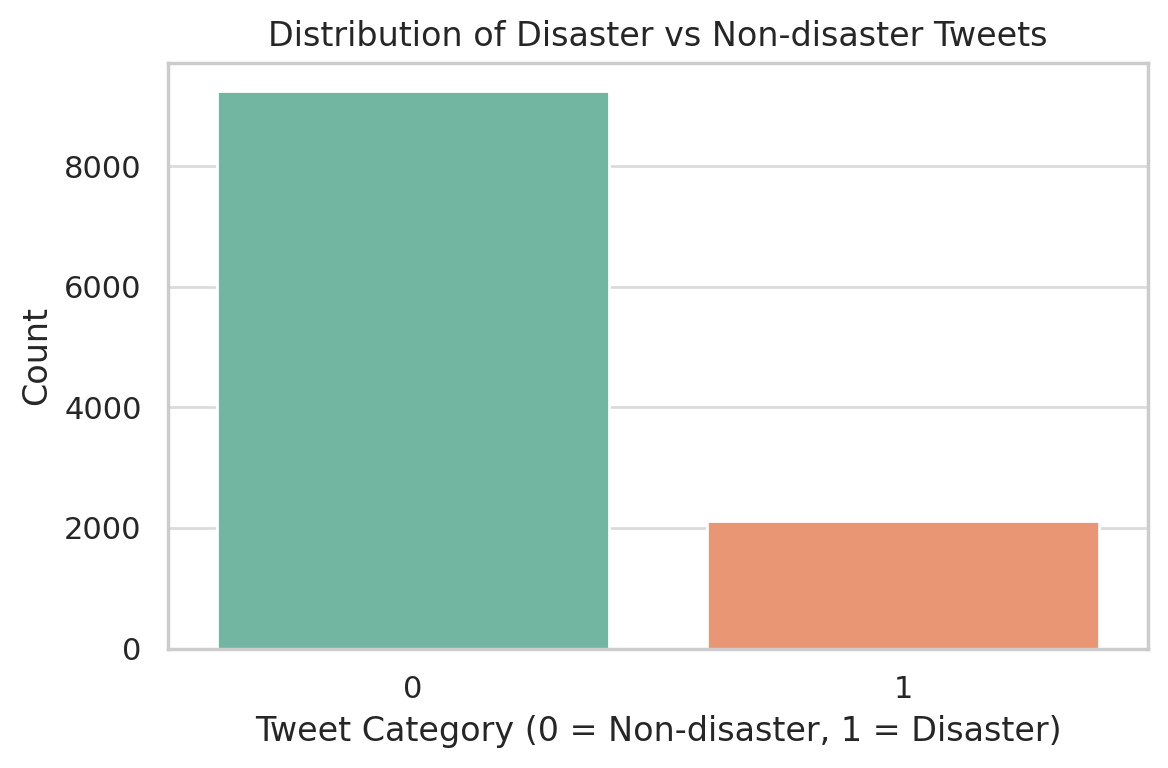
**2. Dataset and Methodology**

**2.1 Dataset**

The dataset consists of 11,370 tweets, which includes fields such as tweet content, keywords, location, and a target label indicating whether the tweet relates to a disaster or not. The **target** field is binary, where:

* 1 represents disaster-related tweets.
* 0 represents non-disaster tweets.

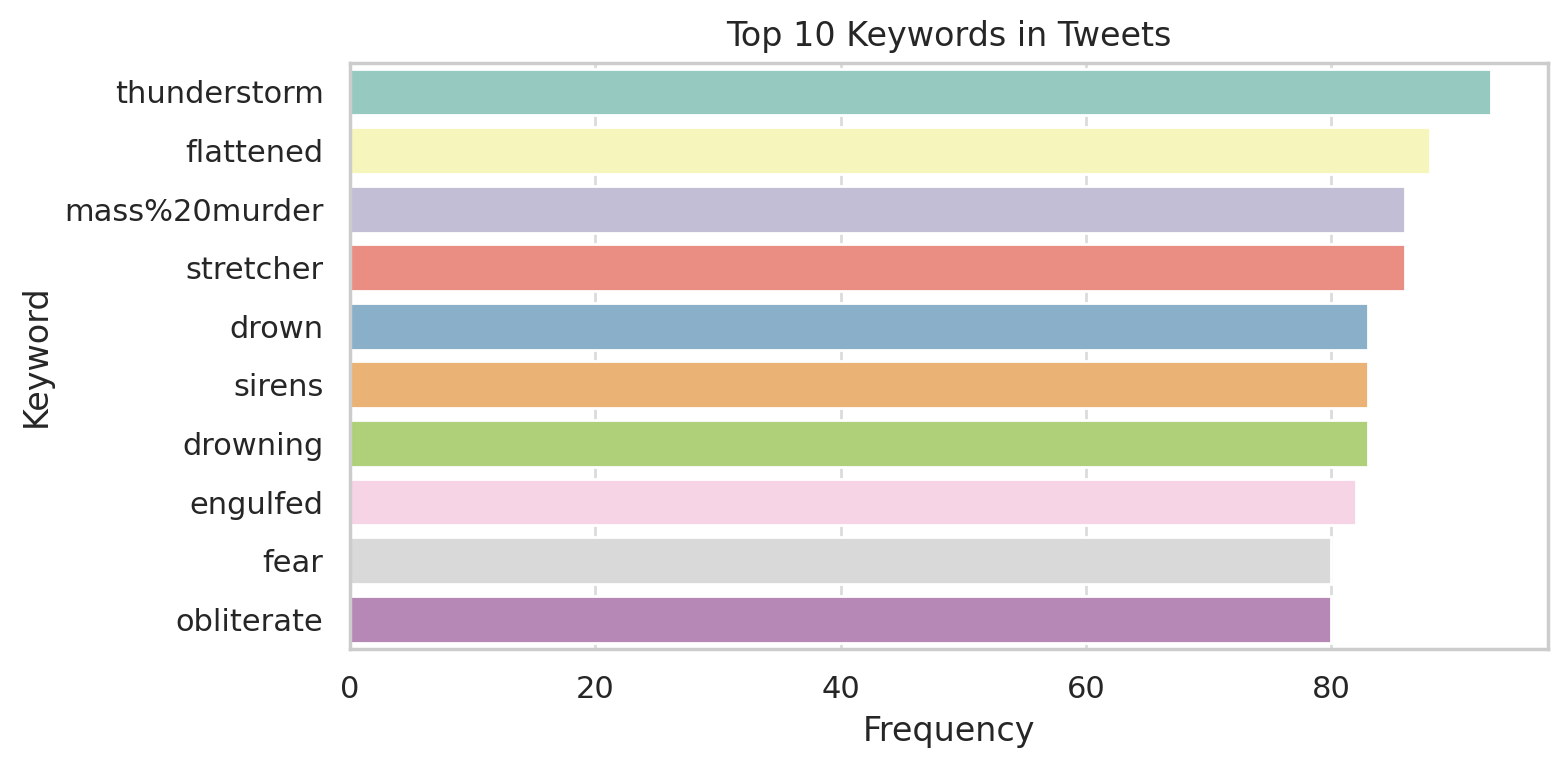
**Figure 1** shows the distribution of disaster vs non-disaster tweets in the dataset.



*Figure 1: Distribution of Disaster vs Non-disaster Tweets*

The dataset also includes a **keyword** field, where each tweet is associated with a keyword (e.g., "ablaze," "earthquake").

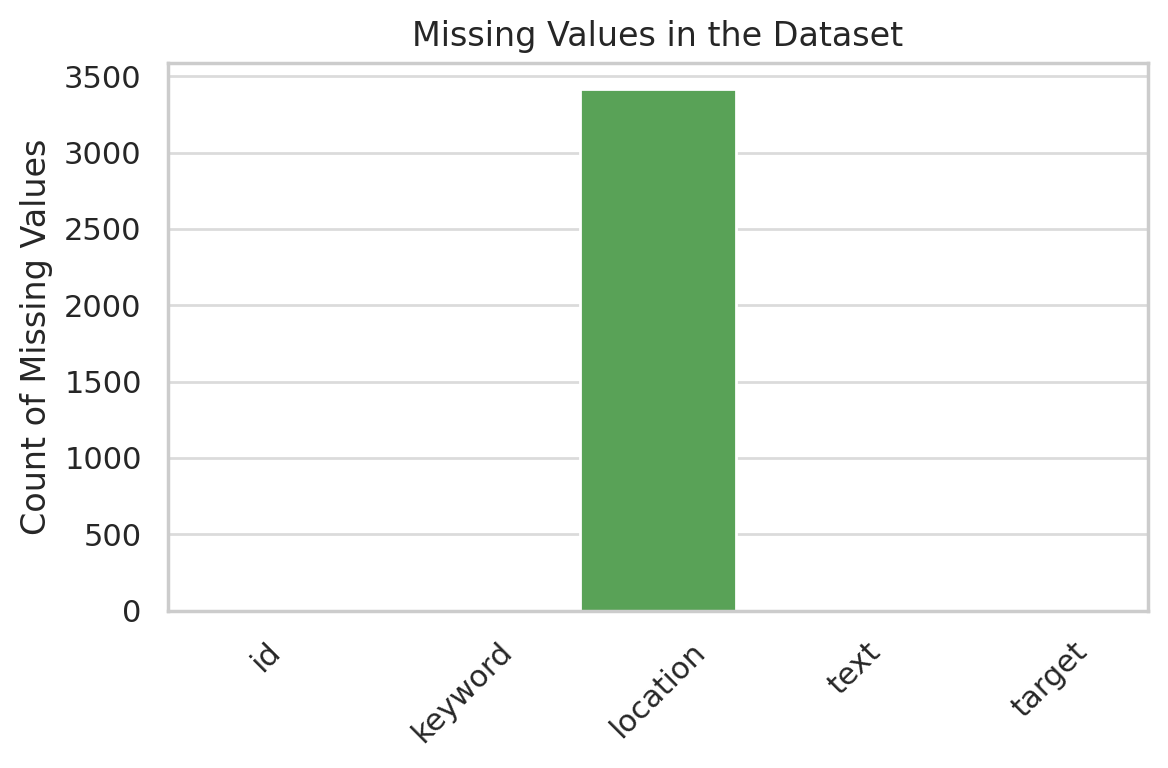
**Figure 2** highlights the top 10 most frequent keywords found in the dataset.



*Figure 2: Top 10 Keywords in Tweets*

There is some missing data, particularly in the **location** field.

**Figure 3** visualizes the missing data across different columns.



*Figure 3: Missing Values in the Dataset*

**2.2 Methodology**

The methodology for this project follows a structured approach, consisting of several key steps:

**2.2.1 Text Preprocessing**

In Week 1, we initiated basic preprocessing tasks:

* **Handling missing values**: Addressing tweets with incomplete data.
* **Basic text cleanup**: Removing special characters, URLs, and irrelevant elements from the tweets.

In Week 2, the preprocessing process was extended to include:

* **Tokenization**: Breaking the tweet text into smaller units (words).
* **Text normalization**: Converting all text to lowercase and eliminating word variations (e.g., stemming, lemmatization).

**2.2.2 Feature Extraction and Representation**

To prepare the text data for machine learning algorithms, we experimented with different feature extraction methods:

* **Bag of Words (BoW)**: Converting tweets into word count vectors.
* **TF-IDF (Term Frequency-Inverse Document Frequency)**: Assigning weights to words based on their importance.
* **Embeddings**: Representing words in a high-dimensional space to capture their semantic meaning.

**2.2.3 Web Scraping**

In addition to working with the existing dataset, the team also experimented with Twitter APIs for web scraping to collect additional disaster-related tweets. Although this is an optional task, it was aimed at expanding the dataset and enhancing model performance.

**3. Results**

**3.1 Week 1**

* **Dataset Confirmation**: The integrity of the dataset was verified, ensuring all records were valid.
* **Initial Exploration**: Preliminary analysis revealed interesting patterns in tweet text length, frequent keywords, and common disaster-related terms.
* **Preprocessing**: Missing values were handled, and the tweet text was cleaned for further analysis.

**3.2 Week 2**

* **Preprocessed Data**: Tweets were successfully tokenized and normalized. The data is now in a structured format for feature extraction.
* **Feature Extraction**: The Bag of Words, TF-IDF, and embedding techniques were applied to the dataset, each producing a different representation of the tweet content.
* **Web Scraping**: Experiments with the Twitter API were conducted, but results have not yet been integrated into the model pipeline.

**4. Conclusion**

The project is on track, with significant progress made in data preprocessing and feature extraction. All deliverables for Week 1 and Week 2 were completed successfully. The preprocessed data is ready for the next phase of classification modeling. Web scraping efforts have begun, and the results will be incorporated in future work.

**5. Future Objectives for the Next Two Weeks**

In the coming two weeks, the following objectives will be pursued:

1. Implementing machine learning classifiers, such as Logistic Regression, Random Forest, and Neural Networks, to classify the tweets.
2. Evaluating the model performance using appropriate metrics like accuracy, precision, recall, and F1-score.
3. Tuning the feature extraction techniques to improve model accuracy.
4. Completing the integration of additional tweets acquired through the Twitter API.

**6. References**

* Dataset Source: [Disaster Tweets on Kaggle](https://www.kaggle.com/datasets/vstepanenko/disaster-tweets)
* Python Libraries: Pandas, Scikit-learn, NLTK, Seaborn, Matplotlib

[1] Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data Preprocessing for Supervised Learning. International Journal of Computer Science, 1(2), 111-117.

[2] Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, 23(4), 3-13.

[3] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

[4] Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Draft Version.