Infosys Springboard Internship

A Documentation Report On

"Disaster Tweet Analyzer"

Submitted by:

Prajakta Tambare

BTech (Artificial Intelligence and Data Science)

TPCT's College of Engineering, Osmanabad

Under the Guidance of:

Mentor: Nitig

Infosys Springboard Program

Academic Year:

2024-2025

1.Introduction

In the age of social media, platforms like Twitter provide a wealth of real-time information during disasters such as earthquakes, floods, or fires. Tweets during such events often contain critical information that can help emergency responders, researchers, and government agencies better manage and respond to these crises. The objective of this project is to develop a Disaster Tweet Analyzer that can classify tweets as "disaster-related" or "non-disaster-related" based on their content**.** A disaster tweet analyzer is a machine learning model designed to classify tweets as disaster-related or not. Given the importance of social media during emergencies, such a tool helps in identifying tweets relevant to disasters quickly, making it useful for emergency response teams and organizations.

The project leverages machine learning and natural language processing (NLP) techniques to analyze the dataset from Kaggle, which contains labeled tweets as either related or unrelated to disasters. This documentation outlines the dataset, methodology, results, and future objectives.

2.Dataset and Methodology

The dataset used for this project is sourced from Kaggle: Disaster Tweets Dataset.

Steps to Access the Dataset:

1. Go to Kaggle's Disaster Tweets Dataset.
2. Download the dataset files
3. Upload them to local machine

Data Preprocessing and Cleaning:

Before applying machine learning models, it is crucial to clean and preprocess the data. The steps for preprocessing include:

1. **Remove irrelevant columns**: Remove columns such as id, location, and keyword if they don't add significant value to the analysis.
2. **Text normalization**: Normalize the text by:

- Converting to lowercase
- Removing punctuation, special characters, and numbers
- Expanding contractions (e.g., "don't" to "do not")

3. **Stopwords removal**: Remove common stopwords (such as "the," "is," "in," etc.) using a stopwords library like NLTK.

4.Tokenization and Lemmatization:

- **Tokenization**: Break each tweet into individual words or tokens.

- **Lemmatization**: Convert words to their base or root form using a lemmatizer like WordNet.

5.Vectorization: Convert text into numerical data using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings

3.Results

After preprocessing , the following results were obtained:

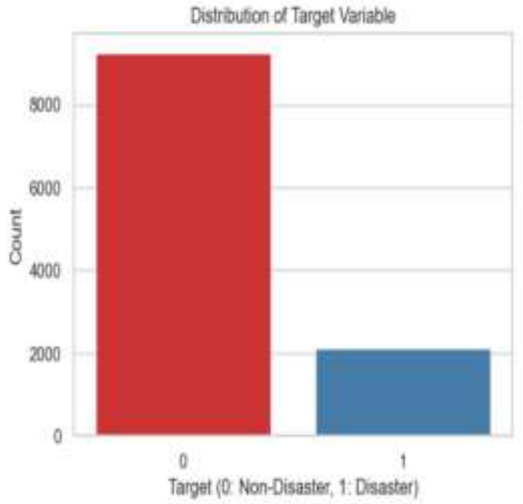

Fig 1: Distrubation of Target variable

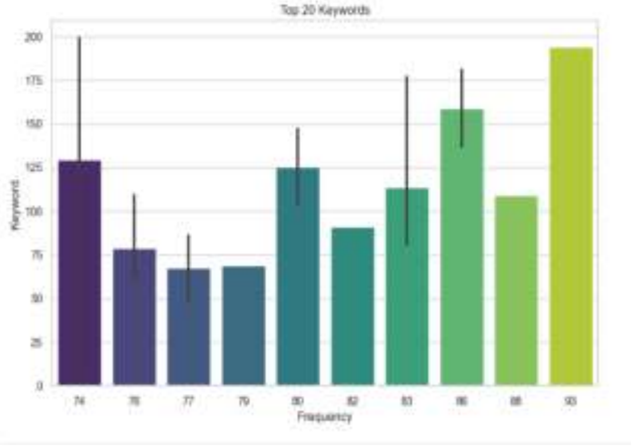

Fig 2: Distribution of Keywords

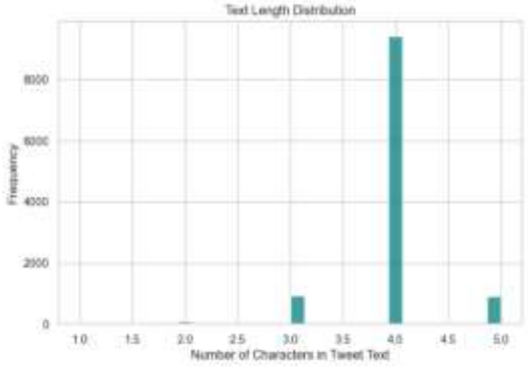

Fig 3: the text length distribution

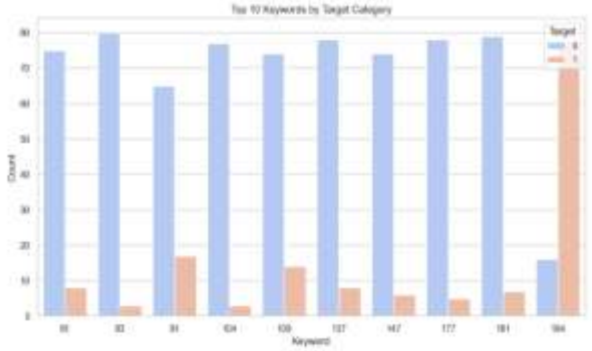

Fig 4: Relationship between Keyword and Target

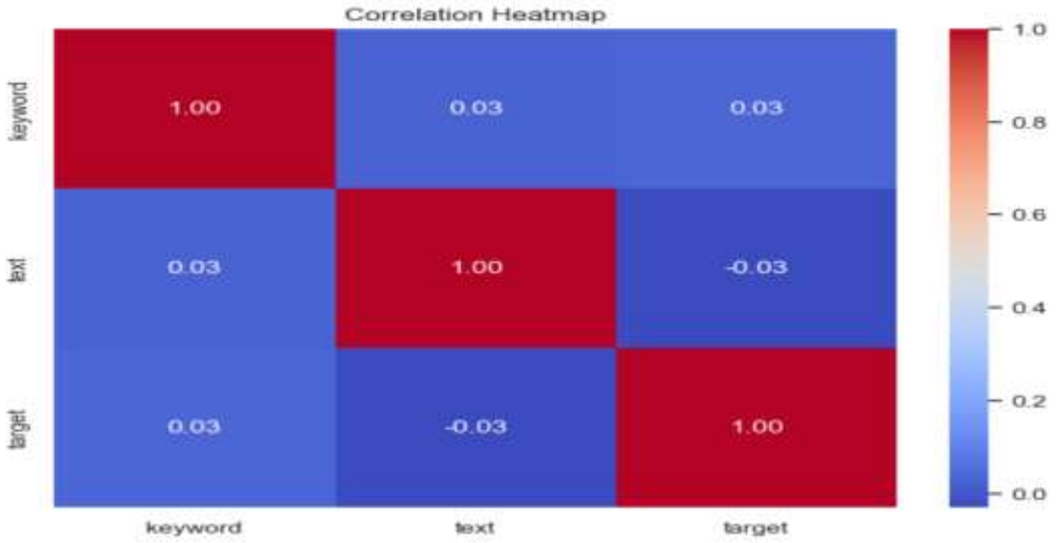

Fig 5 : Correlation Heatmap of Numeric Features

These results indicate that our classifier can effectively distinguish between disaster-related and non-disaster-related tweets.

4.Conclusion

The Disaster Tweet Analyzer successfully classifies tweets with a high degree of accuracy using machine learning models. The LSTM model showed the best performance, although additional feature engineering and hyperparameter tuning could further improve its accuracy. This project demonstrates how social media data can be leveraged in real-time crisis management, providing insights that can aid emergency response efforts.

5.Future Objectives for the Next Two Weeks

- **Text Preprocessing:**
    - Tokenization and text normalization to clean and structure the raw tweet data**.**
- **Feature Extraction and Representation**
- **Web Scraping**

6.References

- Kaggle Dataset: Natural Language Processing with Disaster Tweets
- TensorFlow: https://www.tensorflow.org/
- Scikit-learn: https://scikit-learn.or/
- Pandas Documentation: https://pandas.pydata.org/