Data Collection- *Social Media Integration for Real-Time Disaster Management*

- Marri Bharadwaj (B21CS045)

Data Sources

I used various sources for obtaining the disaster-related tweets as well as normal tweets to build my model on. These sources from which I collected the data are- *CrisisNLP*, *CrisisLex*, *Harvard Dataverse* and also referred to *Kaggle*. These were mainly repositories (collections) of compiled social media posts and tweets, with the first and third sources mainly having disaster related ones (both natural and man-made), and I also included tweets from the other sources to ensure that my model will recognise the difference between disaster and non-disaster related social media posts.

Dataset Description

So, I collected about 7000-8000 social media tweets from the above sources. For diverse applications, I included both natural disasters help tweets (like earthquakes, cyclones, wildfires, floods, etc) and common and uncommon human-made disasters (like industrial accidents, terrorist attacks, oil spills, etc). I compiled these tweets into a *.csv* format. I provided four features to make the model suitable: *text* (social media content), *keyword* (extracted keywords from tweets for easier training), *location* (the site of the disaster as specified by victims or helpers), and *target* (assigned a value of *1* if the post relates to a disaster and *0* if it does not).

Dataset Utilisation and Modelling

- ***Data Pre-processing:*** I will clean the data by removing unnecessary symbols, expanding abbreviations, and standardising location names to ensure consistency.
- ***Feature Extraction:*** I will analyse word frequency and punctuation usage, use word embeddings (FastText), etc., to extract meaningful representations of the tweets.
- ***Model Development:*** Currently, I would implement and train my NLP deep learning models (*LSTM* and *BERT*) to catch long-term dependencies in the textual sequences and for contextual embeddings, including drop-out layers and attention mechanisms.
- ***Assessment and Optimisation:*** Upon training my model as shown above, I would then evaluate model performance with evaluation metrics like accuracy, precision, recall and F1-score; then progressively tune hyperparameters to improve these metrics as needed. Finally, I will deploy my model on additional collected social media posts and show how it works on real-world disaster situations.