**NLP with Disaster tweets**

**Introduction**

Artificial Intelligence (AI) is a technique to help computers think like humans. It is giving computers some of human abilities so that it can help us solve problems intelligently. Image Processing and NLP are few abilities that computers are learning from humans. AI can be used to serve humans beyond their limitations and this is where it’s greatest use lies. They can check large amounts of Data for sensitive information in very less time which can be very difficult for humans.

**Problem Statement.**

NLP with disaster tweets: Twitter is a very popular platform where we can broadcast our thoughts to the entire world bypassing the conventional modes of communication. These tweets can be used to track disasters and emergencies and bring sufficient attention to the concerned authorities.

**Scope of the Project**

It is a case of Supervised learning, Classification Problem. The model will be trained using the twitter dataset. It uses Natural Language Processing & Deep Learning to recognize the intent of the tweet. A working web app which can take a tweet as a link or text and can recognize if it is an emergency. The model will be able to justify its prediction by highlighting the sentences/words on which its judgment is based. The model will be trained in batch processing

**Data Set**

The twitter dataset contains train and test data, with each file having around 11K tweets from which the algorithm learns. These tweets are hand labeled. This data is obtained from Kaggle.

1. id - a unique identifier for each tweet
2. text - the text of the tweet
3. location - the location the tweet was sent from (may be blank)
4. keyword - a particular keyword from the tweet (may be blank)
5. Target (in train.csv only) - this denotes whether a tweet is about a real disaster (1) or not (0)

**Methodology**

Data preprocessing:

* **Tokenization:** is splitting text to meaningful chunks.
* **Stemming/Lemmatization**: remove word suffixes to obtain the root word.
* **Hashing/Vectorizing:** maps the given word to a unique key for faster access.
* **Sparse Matrix Encoding** (According to requirement)
* **Dimensionality reduction:** NPL uses large dimensional sparse matrices, it makes sense to reduce dimensions with a little reduction in accuracy.

Text classification

* Fitting and training the model using the above generated data.
* Training various NLP and Deep Learning models to predict.
* To classify if the input tweet is really emergency

**Deliverables**

Customizable model with default setting. These settings can be changed to select the desired training algorithm. The output will be if the tweet means emergency with a confidence interval.

P.s. Beyond Scope

A full project would involve streaming the tweets continuously and recognising the tweets which communicate actual disaster. Then the algorithm would respond by sending notifications to authorities nearest to the place of the event. This algorithm is expected to work within a certain confidence interval. The model needs computational resources and storage space to store its data. It is a Limited memory model where it learns from the incoming data but need not store much of the data.