**Disaster Tweets Classification**

Members:

* Lohith Reddy Ramireddy 200001066
* Sapavath Shashi Kumar 190001055

**Problem statement:**

* Twitter is an emergency communication channel in times of emergency.
* Since use of smartphones has increased drastically these days, which enables people to announce emergency immediately
* Because of this more agencies are interested to programmatically monitoring twitter.
* A tweet might be clear to human that they are in emergency, but may not be clear for a machine.
* **So we build a machine learning model that predicts which tweets are real about disasters and which are not.**
* It’s basically a **binary classification model** into REAL or NOT.

**Data set collection:**

* In this study, I’ve used the social media disaster tweets data set provided by Kaggle.
* Its freely available to view and download on Kaggle.com.
* In this dataset, they’ve provided more than 10,000 tweets.
* We estimate whether a tweet is regarding disaster or not from these 10,000 tweets.
* Each entry in the data set includes the following details:

1. Id – tweet No.
2. Text – text of the tweet.
3. Location – location from which tweet is delivered.
4. Keyword – a keyword like bushfire, forest fire, cyclone etc.
5. Target – whether the tweet is REAL disaster or NOT.

**Data pre-processing:**

* Data pre processing has different parts in it which includes:

1. Data Cleaning
2. Changing text to all lower or all upper
3. Removing noise
4. Tokenization
5. Stop word removal
6. Stemming
7. Lemmatization

**Approach:**

* Here I’ll be using some already developed models to classify the tweets into REAL disasters or NOT.
* Here I’ll be implementing 3 classification models and use the model which gives me more accurate predictions.

1. Logistic Regression model
2. Decision tree classifier model
3. Random forest classifier model

**Model performance metrics:**

* Here ill be using these 3 performance metrics:

1. Confusion matrix
2. F1 score
3. Accuracy

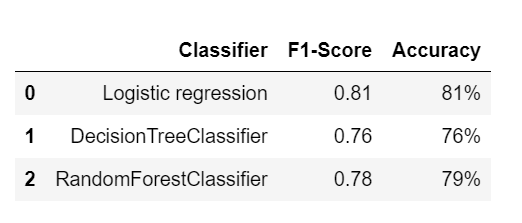
**Vectorization of text:**

* There are mainly 2 types of vectorization we use:

1. Count vectorization
2. Tf-idf vectorization

**Results:**

For Count Vectorization:



For Tf-IDF Vectorization:

