**Disaster Tweet Analyzer: Using NLP**

**Task-1 REPORT**

**Submitted to Mentor - Nitig Singh Sir**

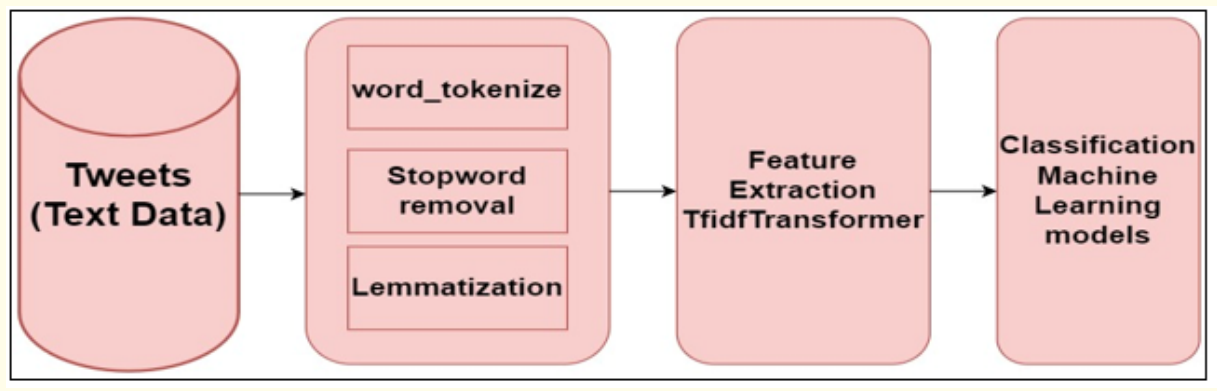
**By- Ashutosh Kumar Yadav**

* **Introduction**

In the realm of disaster management, social media platforms like Twitter have emerged as valuable sources for real-time information dissemination and situational awareness. The project at hand aims to develop a machine learning model that analyzes tweets to predict whether a tweet is genuinely related to an ongoing disaster or merely discussing a related topic or whether the tweet is fake of giving wrong information. By distinguishing between actionable information and general chatter, this model can aid authorities and first responders in identifying relevant tweets for timely decision-making. The model will leverage natural language processing (NLP) techniques [2] and machine learning (ML) algorithms to automatically classify tweets, improving the efficiency of information extraction during disaster scenarios.

The development process will involve data collection, preprocessing, and feature extraction using various NLP techniques such as tokenization, stopword removal, and vectorization [2]. We will train different ML models, including classical algorithms and modern deep learning approaches, to determine which performs best for this task. The objectives of this disaster tweet analyzer are threefold: (1) to accurately classify tweets as disaster-related or not, (2) to enhance the speed and quality of disaster response by filtering relevant information, and (3) to establish a system that can adapt and improve over time with new data, ensuring continued relevance and effectiveness.

* **Dataset and Methodology (Exploration)**

****

**Image showcasing the process of model building [2]**

For this project, datasets were chosen from Kaggle.com after careful consideration of various available options. The primary criterion for selection was the number of tweets in the dataset, as a larger dataset would provide more data for training and evaluating the machine learning model. Additionally, the type of data played a crucial role; since the model aims to perform text-based analysis, datasets containing clean, well-annotated text data were prioritized. After reviewing multiple datasets, two were selected based on these parameters for their comprehensive coverage of disaster-related tweets and clear labelling of whether a tweet was relevant to a real disaster or not [1].

The two selected datasets were then explored to understand their structure and characteristics. This exploration involved examining the number of samples, types of disaster labels available, and distribution of disaster-related versus non-disaster-related tweets. The datasets contained various columns, including tweet text, labels indicating relevance to a disaster, and additional metadata. These datasets provided a good foundation for building a robust disaster tweet analyzer, ensuring diverse examples and varied scenarios to train and test the model effectively.

Finally, this dataset for finalised for further exploration. [Dataset](https://www.kaggle.com/datasets/vstepanenko/disaster-tweets/data)

Some basic info about dataset is following-

The file contains over 11,000 tweets associated with disaster keywords like “crash”, quarantine”, and “bush fires” as well as the location and keyword itself. The data structure was inherited from [Disasters on social media](https://appen.com/resources/datasets/)

The tweets were collected on Jan 14th, 2020.

Some of the topics people were tweeting:

* The eruption of Taal Volcano in Batangas, Philippines
* Coronavirus
* Bushfires in Australia
* Iran downing of the airplane flight PS752

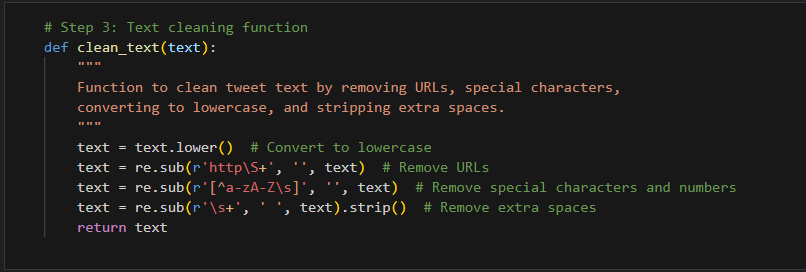
The tweets are not structured and need to be pre-processed. So to perform pre-processing the data and to make it ready for analysis we need to clean and process the data. We’ll be working on the dataset and its processing in jupyter notebook.

So, the step 1 is to load the dataset into our Jupyter notebook.

Here we observed a technical problem, which stated- binary incompatibility issue between pandas and numpy. This usually happens when the installed versions of pandas and numpy are not compatible with each other.

Solution- we just updated tour packages before uninstalling it and installing it again using command prompt.

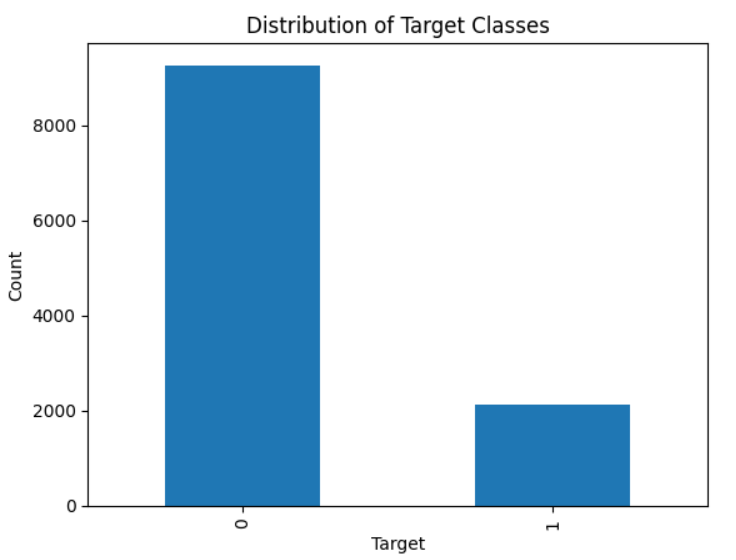
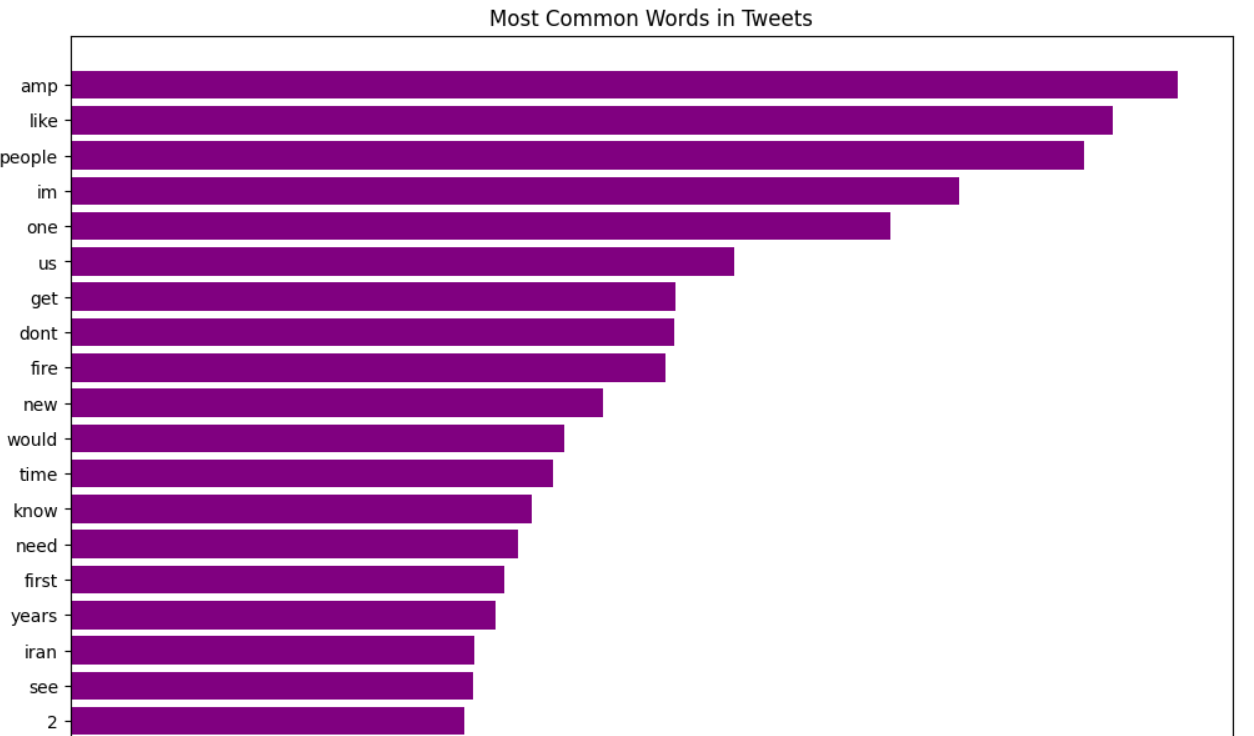
Now importing the required packages and just applying the data cleaning steps like handling missing values, cleaning tweets like removing char, urls, also converting ever fond to lowercase.



After cleaning the required dataset, the cleaned dataset is saved for further cleansing.

* **Results**

On further data visualisation some pointers are drawn in form of graphs and charts.

** **

**Textual Data Analysis**

****

**Target text v/s Normal text clouds**

* **Conclusion**

With the dataset selection and exploration completed, the project has successfully moved through its initial phases. The chosen datasets were loaded into a Jupyter Notebook, where data cleaning steps such as removing special characters, punctuation, and unwanted symbols were performed to prepare the text data for further processing. Data visualization techniques were applied to better understand the distribution of disaster and non-disaster tweets, providing insights into the dataset's characteristics. The visualizations, included in this report, highlight key patterns that will guide the next steps. Moving forward, tokenization, stopword removal, and lemmatization will be performed to fully preprocess the text for model training.

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

Over the next two weeks, the focus will be on advancing the text preprocessing and analysis phases to prepare the dataset for building a robust machine learning model. The upcoming tasks will include tokenization, lemmatization, and stemming, which will help standardize the text data by breaking it into individual words, reducing words to their root forms, and eliminating variations. Following this, exploratory data analysis (EDA) will be conducted to uncover meaningful patterns, relationships, and insights from the data. Additionally, feature engineering will be carried out to create relevant features that can enhance the model's performance. These steps will collectively lay the groundwork for effective model training and accurate prediction of disaster-related tweets.

* **References**

1. Shekhar, H., & Setty, S. (2015, August). Disaster analysis through tweets. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1719-1723). IEEE.
2. Lakshmi, S. D., & Velmurugan, T. (2023). Classification of disaster tweets using natural language processing pipeline. *Acta Sci Comput Sci*, *5*(3).
3. Apampa, A. D., & Li, N. (2022). Natural Language Processing for Disaster Tweets.