Infosys Internship 5.0

Disaster Tweet Analysis with Natural Language Processing

BY– Moksha Kankariya .

**Introduction :-**

Disaster-related tweets have the potential to alert relevant authorities early on so that they can take action to reduce damage and possibly save lives. The **Disaster Twitter Analyzer** project focuses on using Natural Language Processing (NLP) and Artificial Intelligence to analyze tweets related to disasters. The objective of the project is to detect and classify tweets as either related to a disaster (such as earthquakes, floods, or wildfires) or unrelated, and possibly further categorize the disaster-related tweets into specific types of disasters. This can be particularly useful for real-time disaster response and awareness.

**Dataset and Methodology (Exploration) :-**

For the **Disaster Twitter Analyzer** project, I have used dataset that contains tweets labeled as disaster-related or non-disaster-related. One popular dataset for this purpose is the **Kaggle Disaster Tweets dataset**. The [training dataset](https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets#Training_data_set) contains information on 7613 tweets, each with a unique id, keyword (if available), location (if available), text and whether or not the tweet indicates a real disaster or not (expressed via a binary variable).

**Disaster-related**: Tweets that provide information about natural disasters, accidents, or other calamities.

**Non-disaster-related**: Tweets that are not relevant to disasters, like daily social interactions.

The features of the dataset are described below.

id : A unique identifier corresponding to the tweet.

keyword : A highlighting word from the tweet.

location : The location from where the tweet is sent.

text: The textual content of the tweet.

target : A binary variable, which is 0 if the tweet does not indicate a real disaster and 1 if it does.

**Methodology:-**

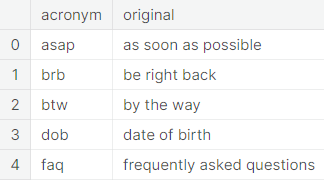
**1) Text Normalization**

**Text Cleaning**: Tweets often contain noisy elements such as URLs, mentions, hashtags, emojis, and special characters. Cleaning involves:

Removing URLs, hashtags, mentions (@username), and punctuations.Converting text to lowercase.

[**Substitution of acronyms**](https://sugatagh.github.io/dsml/projects/natural-language-processing-with-disaster-tweets/#substitution-of-acronyms)**:**

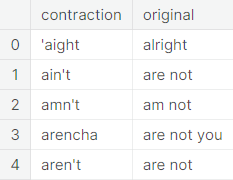
Acronyms are shortened forms of phrases, generally found in informal writings such as personal messages. Examples:



These time and effort-saving acronyms have received almost universal acceptance in social media platforms including twitter. For the sake of proper modeling, we convert the acronyms, appearing in the tweets, back to their respective original forms.

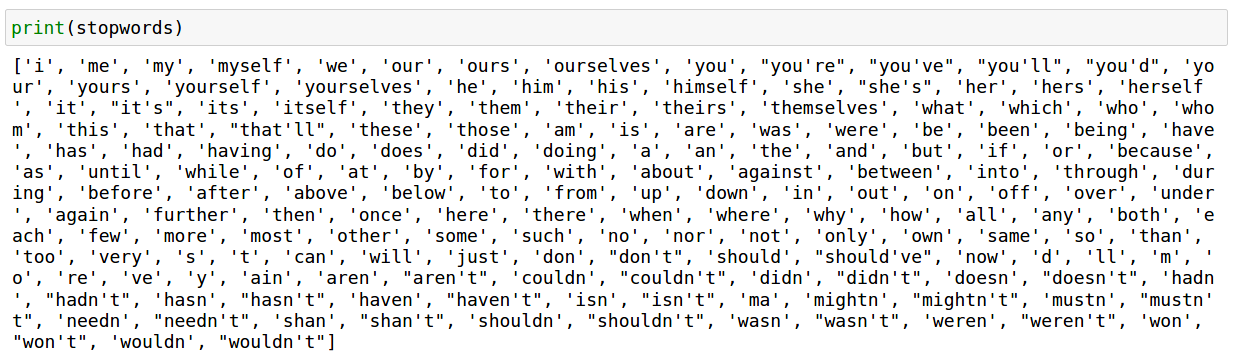
**Substitution of contractions:**

A contraction is a shortened form of a word or a phrase, obtained by dropping one or more letters. Examples:



**Removal of stop words**

Several words, primarily pronouns, prepositions, modal verbs etc., are identified not to have much effect on the classification procedure. To get rid of the unwanted contamination effect, we remove these words. For this purpose, we use the stopwords module from [NLTK](https://en.wikipedia.org/wiki/Natural_Language_Toolkit). Some of these words are shown below.



**Stemming and lemmatization :**

Stemming is the process of reducing the words to their root form or stem. It reduces related words to the same stem even if the stem is not a dictionary word. For example, the words introducing, introduced, introduction reduce to a common word introduce.

Lemmatization offers a more sophisticated approach by utilizing a [corpus](https://en.wikipedia.org/wiki/Text_corpus) to match root forms of the words. Unlike stemming, it uses the context in which a word is being used. Upon applying lemmatization, the same sentence becomes introduce lemmatization as an improvement over stem. Here we use the [spaCy](https://en.wikipedia.org/wiki/SpaCy) lemmatizer.

**Result :-**

The results of the **Disaster Twitter Analyzer** project will largely depend on the quality of the dataset, model selection, and the overall methodology applied.

**Conclusion :-**

The **Disaster Twitter Analyzer** project demonstrates the potential of using Natural Language Processing (NLP) and Artificial Intelligence (AI) to automatically classify and monitor disaster-related tweets in real-time. The project successfully combines various machine learning and deep learning techniques to process raw tweet data, clean and transform it into useful features, and classify the tweets into disaster-related and non-disaster-related categories. With real-time tweet classification, this system can provide an early warning mechanism, helping authorities react quickly to emerging disasters by gathering information directly from people on the ground.

**References :-**

https://www.kaggle.com/code/sugataghosh/natural-language-processing-with-disaster-tweets

<https://github.com/levist7/NLP_Disaster_Tweet_Analysis>

<https://github.com/sugatagh/Natural-Language-Processing-with-Disaster-Tweets>