Moksha Jitendra Kankariya

Disaster Tweet Analysis with Natural Language Processing

Infosys Internship 5.0

**Introduction :-**

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**. It contains 10,000 tweets classified into two categories:

**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.

**Methodology:-**

**1)Data Preprocessing**-

**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.

Removing stopwords (common words that don’t add much meaning to text, like "and", "the", etc.).

Lemmatization or stemming (reducing words to their base form).

**Tokenization**: Splitting each tweet into individual words (tokens) to analyze them.

**Handling Imbalanced Data**: If disaster-related tweets are less common, using techniques like oversampling the minority class or undersampling the majority class may be necessary.

**2)Feature Extraction**-

**Word Embeddings**: Using pre-trained embeddings like **bagOfWords**, **GloVe**, or **BERT** to capture the semantic meaning of words in the tweets.

**Model Selection**: Several machine learning and deep learning models can be used for text classification:

**3)Traditional Machine Learning Models**:

**Logistic Regression**: Works well for binary classification tasks with linear decision boundaries.

**Random Forest**: An ensemble method that builds multiple decision trees to improve accuracy.

**Deep Learning Models**:

**Recurrent Neural Networks (RNN)**: Used for processing sequences of data, such as text. **LSTM (Long Short-Term Memory)** is a variant of RNN that captures long-term dependencies in text.

**Convolutional Neural Networks (CNN)**: Surprisingly effective for text classification, CNN can be applied by treating text data as a matrix.

**4)Model Training**:

Split the data into **training** and **testing** sets (e.g., 80% training and 20% testing).

Use cross-validation techniques (like k-fold cross-validation) to evaluate model performance on multiple splits of the data.

Train the model on the training data and tune hyperparameters (e.g., learning rate, batch size, number of epochs) for best performance.

**5)Real-time Tweet Classification**:

Use the trained model to classify real-time tweets fetched from the Twitter API.

Track keywords related to disasters (e.g., earthquake, fire, flood, etc.) and filter tweets using location-based filtering if necessary.

Once tweets are classified as disaster-related, they can be further analyzed and visualized on a dashboard.

**6)Visualization and Reporting**:

**Data Visualization**: Plot trends in disaster-related tweets over time. For instance, a spike in earthquake-related tweets can be visualized on a timeline.

**Geolocation Mapping**: Map tweets to their respective locations if geolocation data is available, helping authorities identify disaster hotspots.

**Dashboard**: Create a user-friendly dashboard for tracking disaster-related tweets in real-time, displaying metrics, counts, and categories of disasters.

**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