DISASTER TWEET ANALYZER

Project Documentation

1. Introduction

Twitter has become an essential and effective medium to disseminate information, especially in emergencies. Due to this fact, many people and organizations monitor Twitter activity to locate tweets about certain significant events, like weather emergencies. This project aims to use machine learning to determine which tweets are about natural disasters and which are not.

By leveraging Natural Language Processing (NLP) techniques, this project focuses on building a machine learning model to automatically detect disaster-related tweets and verify them. The end goal is to filter and classify relevant information, contributing to real-time situational awareness during natural or man-made disasters.

2. Dataset and Methodology

Dataset:

The dataset used for this project is the "**Kaggle's NLP Disaster Tweet Classification Dataset",** which contains 10,873 tweets, each labelled as either **disaster-related** or **non-disaster-related**. The variables are:

- **Id**: A unique integer assigned to each tweet.
- **Location:** the location from where the tweets were sent.
- **Text:** The list of sample tweets.
- **Target:** A list of predictions for the tweets, the model constructed is to match these predicted values as closely as possible. Since this is the data which we are using to train our model, we will be using the tweets to formulate our model.

Data Preprocessing:

Before feeding the data into the model, we apply several preprocessing steps:

1. **Removing URLs and Special Characters**: Cleaning up tweet text by eliminating noise like URLs, hashtags, and unnecessary symbols.

2. **Lowercasing**: Ensuring that all the text is in lowercase for uniformity.
3. **Tokenization**: Splitting the tweet text into individual words (tokens).
4. **Removing Stop Words**: Filtering out common words that don't contribute to the meaning of the sentence (e.g., "is", "and", "the").

Methodology & Proposed System:

The system is composed of three different modules: Extraction of Data, Sorting of Data and Analysis of Data. Each module is further divided in submodules which cater to specific tasks. Figure 1 gives out the high-level system diagram of the application. The different phases of the system are detailed down as follows:

A. Data Extraction:

This module deals with the extraction of tweets from the web and storing them locally for analysis.

1) Tweet Capturing*:* Data extraction was mainly done with the help of Twitter Streaming API. This API is capable of capturing live tweets from the web and parsing them as String objects. These String objects can be further worked upon by the program for storage and manipulation.

2) Redundancy Check: It is possible for the program to capture the same tweet more than once due to re-tweets by the user, recapturing of older tweets etc.

3) Data Storage: The tweets which pass the Redundancy Check sub-module are then stored systematically in files in ascending order of their time of posting. The system stores the tweet message content, the username of the person who posted the tweets and time-stamp of the tweet. The system is capable of storing even more data such as number of re-tweets and comments upon the tweet.

B. Data Sorting: The stored tweets are now analysed and their disaster category is determined. This is done by checking each tweet, against a set of predefined weighed keywords. A tweet is said to be belonging to a particular category of disasters if it matches 40 percent of the listed keywords. Matching of higher weighed keywords provides greater relevance factor to a tweet. The tweets are then stored in then separate files pertaining to their disaster category along with their relevance factor. The tweets which do not belong to any category are readily discarded.

C. Analysis of Data: The categorized data is then passed on to the third module for analysis and interpretation. This module is divided into four distinct sub-modules which check the data for Disaster Distribution, Geo-Tagging, Occurrence Frequency and Sentiment Rating.

1) Geo-Tagging*:* The location of disaster is the foremost matter of interest in any type of disaster analysis. To extract the location from the tweets, a geo-filter tag is applied on the tweets to determine their point of origination. However, in many cases tweets are posted by people who are not actually tweeting from the affected region. Often in such a case the disaster location is mentioned within the tweet. The location mentioned in the tweet can be determined by the help of Google Maps JavaScript API wherein all the tweets are passed through a method one by one which splits the tweets into words and tries to find the co-ordinates of that word through the said API.

2) Disaster Distribution Analysis: This module categorizes the distribution of various types of disasters in the world. It makes use of location determined from the previous submodule and clusters the tweets based on location co-ordinates mentioned against tweets using K-Nearest Neighbour Algorithm.

3) Disaster Occurrence Frequency: Each type of disaster data is analysed by monitoring the tweet incoming rate for different locations via the geo-tagging module. This live data is aggregated into clusters and different cluster sizes are then analysed for a particular category of disasters. The number of people tweeting about a particular disaster within a given time period is taken as an indicator for the disaster frequency.

Results:

This model still has room to improve and it could be refined in the future to improve its accuracy.

Conclusion:

The project demonstrated the effectiveness of NLP techniques in disaster tweet analysis. Using a combination of data preprocessing, TF-IDF vectorization, and machine learning models.

5. Future Objectives for the Next Two Weeks

The immediate goals for the next two weeks are:

1. **Improve Model Accuracy:**
   - Experiment with more advanced transformers such as **BERT (Bidirectional Encoder Representations from Transformers)** for better understanding of tweet context and semantics.
   - Fine-tune hyperparameters of existing models like Random Forest and LSTM.
2. **Deploy the Model:**
   - Set up a simple **web interface** where users can input tweets or hashtags, and the system will classify tweets as disaster-related or not.
   - Use **Flask/Django** for creating the web API and **Heroku** for deployment.
3. **Real-time Tweet Stream Integration:**
   - Implement **Twitter API** to fetch live tweets, allowing the model to classify real-time data streams.
4. **Sentiment Analysis Integration:**
   - Add a **sentiment analysis** module to understand the emotional tone of the tweets, which can further guide response strategies.

6. REFERENCES

[1] S. B. Liu, L. Palen, J. Sutton, A. L. Hughes, and S. Vieweg, "In search of the bigger picture: The emergent role of on-line photo sharing in times of disaster," in *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM)*, 2008.

[2] L. Palen and S. Vieweg, "The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*.

ACM, 2008, pp. 117–126.

[3] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, “Chatter on the red: what hazards threat reveals about the social life of microblogged information,” in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 2010, pp. 241–250.

[4] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, “Microblogging during two natural hazards events: what twitter may contribute to situational awareness,” in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 1079–1088.

[5] L. Palen, S. Vieweg, S. B. Liu, and A. L. Hughes, “Crisis in a networked world features of computer-mediated communication in the April 16, 2007, Virginia tech event,” *Social Science Computer Review*, vol. 27, no. 4, pp. 467–480, 2009.

[6] J. Sutton, L. Palen, and I. Shklovsky, “Backchannels on the front lines: Emergent uses of social media in the 2007 southern California wildfires,” in *Proceedings of the 5th International ISCRAM Conference*. Washington, DC, 2008, pp. 624–632.

[7] S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu, “Tweet tracker: An analysis tool for humanitarian and disaster relief.” in *ICWSM*, 2011.

[8] R. E. Cohn and W. A. Wallace, *The Role of Emotion in Organizational Response to a Disaster: An Ethnographic Analysis of Videotapes of the Exxon Valdez Accident*. Natural Hazards Research and Applications Information Centre, Institute of Behavioural Science, University of Colorado, 1992.

[9] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, and J. Rodrigue, “A demographic analysis of online sentiment during hurricane

irene," in *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 2012, pp. 27–36.

[10] J. B. Lee, M. Yba˜nez, M. M. De Leon, and M. R. E. Estuar, "Understanding the behaviour of Filipino twitter users during disaster," *Journal on Computing (JoC)*, vol. 3, no. 2, 2014.

2015