Disaster Tweet Analyzer-document for week 1 and 2:

Submitted by :Sajja Yuva Sai Rithvik

Submitted to:Nitig Singh

**Introduction**

Over the past two weeks, I've been working on the Disaster Tweet Analyzer project. This initiative aims to use Natural Language Processing (NLP) to identify critical information in tweets during natural disasters. I've focused on developing an NLP model that can distinguish between disaster-related and irrelevant tweets.

My goal is to create a tool that will help emergency responders, government agencies, and the public quickly access important information during crises. I've made significant progress in defining project objectives, implementing initial text preprocessing steps, and exploring strategies for web scraping additional tweets from Twitter.

**Dataset and Methodology**

For the project, the primary dataset derived was taken from Kaggle's "Disaster Tweets" dataset consisting of 10,000 labelled tweets. Every tweet has been classified as either related to a disaster with a value of 1 or not with a value of 0. The structure and intricacies of the dataset were explored in detail in the first phase.

**Text Preprocessing**

I developed a preprocessing pipeline to clean up the noisy tweet data:

* I implemented tokenization to break down each tweet into individual words.
* I wrote a function to convert all text to lowercase for consistency.
* I created a script to remove punctuation, URLs, hashtags, and mentions. I'm considering reintroducing disaster-specific hashtags later.
* I expanded contractions (e.g., "can't" to "cannot") and filtered out stop words.
* I applied lemmatization to reduce words to their root forms.

**Feature Extraction and Representation**

After cleaning the tweets, I tested different feature extraction methods:

I implemented a Bag of Words (BoW) model to count word frequencies.

I applied Term Frequency-Inverse Document Frequency (TF-IDF) for a more nuanced word importance weighting.

I explored Word Embeddings techniques like Word2Vec and GloVe to capture semantic relationships.

**Results**

Through my exploration of the dataset, I found an even distribution between disaster-related and non-disaster tweets. My initial analysis revealed that disaster-related tweets tend to be more detailed and often contain critical keywords like "fire," "earthquake," and "help."

I successfully applied my text preprocessing steps, including tokenization and lowercasing, to prepare the dataset for further analysis. I created word clouds to visualize the most frequent words in both categories of tweets. These early findings have given me valuable insights into the patterns of disaster-related language, which will be crucial for the next stages of the project.

**Conclusion**

My initial exploration and text preprocessing have highlighted key differences between disaster-related tweets and other tweets. I identified urgent keywords and more descriptive language as distinguishing features of disaster-related content. By removing noise and standardizing the text, I've prepared the dataset for more advanced analysis.

The steps I've taken, such as tokenization and feature extraction trials, have positioned the project well for the next phase of classification model development. My early results, particularly in identifying common disaster-related keywords, suggest that this approach will be effective in distinguishing between tweet categories.

**Future Objectives**

I plan to implement more sophisticated feature extraction techniques, such as n-grams, to capture word sequences and their contextual meaning.

I will experiment with various machine learning models, including Random Forest and Support Vector Machines (SVM). I'm particularly interested in exploring deep learning approaches like Long Short-Term Memory (LSTM) networks.

I aim to improve the web scraping pipeline to handle real-time data more effectively. This will include refining my techniques for removing emojis and applying sentiment analysis.

I will evaluate the performance of my models using precision, recall, and F1 score. I plan to use hyperparameter tuning and cross-validation to optimize model accuracy.

Finally, I intend to start developing a user-friendly interface to display tweet classifications intuitively.

It would create an interface that is user-friendly and makes all classifications of the tweets available in a way that users can access insights about them in an intuitive manner.

**References**

 Kanojia, G. & Rastogi, A. *Disaster Tweets Classification*.

 Sharma, A., Thakur, K., Kapoor, D. S., Singh, K. J., Saroch, T., & Kumar, R. *Disaster Analysis Through Tweets*.