Infosys Springboard Internship 5.0

Disaster Tweet Analyzer: Using NLP

Submitted to Mentor - Nitig Singh Sir

By- Shubham Gohil

Introduction

Social media, particularly platforms like Twitter, has become a vital source of real-time information during disasters, helping both emergency response teams and those affected. However, the sheer volume of tweets shared during such events makes it difficult to sift through and identify important updates. This project aims to solve this problem by developing a model that automatically classifies tweets as either disaster-related or not, utilizing natural language processing (NLP). This will enable responders to prioritize essential information, enhancing coordination and decision-making during emergencies. The difficulty lies in the brevity and ambiguity of tweets, where similar wording can describe both disaster and non-disaster situations.

The project applies established text classification methods, such as Bag of Words, TF-IDF, and deep learning, to better understand the language used in disaster-related tweets. At this stage, the focus is on cleaning the data, identifying common keywords in tweets about disasters, and preparing for model training. By examining the most frequently used terms in disaster-related posts, the project will develop a more accurate model for classifying tweets. In the long term, the aim is to improve real-time disaster response by providing authorities with the tools to quickly identify and react to critical situations.

Dataset and Methodology

The dataset used in this project is a collection of tweets, each labeled as either disaster-related or not. The main attributes of the dataset include:

- id: A unique identifier for each tweet.
- text: The tweet content, which is the primary input for the analysis.
- target: A binary label indicating whether the tweet is disaster-related (1 for disaster-related, 0 for not disaster-related).

The methodology can be broken down into several key steps that form the foundation for the tweet classification process:

1. **Data Preprocessing**: The first crucial step is to clean and prepare the text data so it can be effectively analyzed. In this notebook, several preprocessing steps are involved:
    - **Tokenization**: This involves breaking each tweet down into individual words, known as tokens. Tokenization helps in treating each word separately for analysis, making it easier for the machine learning model to understand the structure of the text.
    - **Stopword Removal**: Common words like "the," "and," and "is" appear frequently but carry little useful meaning in classification. Removing these stopwords reduces noise in the data and allows the model to focus on more informative terms.
    - **Stemming or Lemmatization**: These processes reduce words to their base or root forms. For example, “running” might be reduced to “run” so that different forms of the same word are treated as equivalent, ensuring consistency across the dataset.
    - **Removal of Punctuation and Links**: Tweets often contain punctuation marks, URLs, or other symbols that don’t provide any useful information. Removing these elements ensures that the model is only focused on the core semantic content of the tweet.
2. **Exploratory Data Analysis (EDA)**: After preprocessing, the next step is to explore the data to gain insights into the structure and content of the disaster-related tweets. EDA typically involves:
    - **Word Frequency Analysis**: This involves counting how often each word appears in the disaster-related tweets. By analyzing word frequency, we can identify the most common terms used when people discuss disasters. For example, words like “flood,” “fire,” “help,” and “earthquake” may appear frequently, providing clues about the types of disasters discussed.
    - **Data Visualization**: The results of the word frequency analysis are often presented visually, such as through bar charts or word clouds. These visuals highlight the most frequent words in disaster-related tweets, making it easier to understand which terms are most important for classification. By identifying common keywords, we gain valuable insights into the language typically used in disaster situations, which can help improve the model’s accuracy in distinguishing relevant tweets.

This methodological approach ensures that the data is properly cleaned and understood before model training, laying the groundwork for building a robust classification model that can efficiently detect disaster-related tweets.

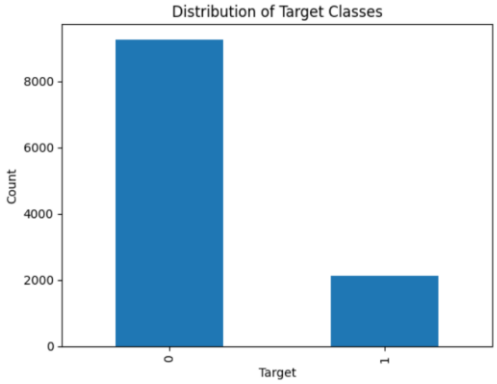
Results

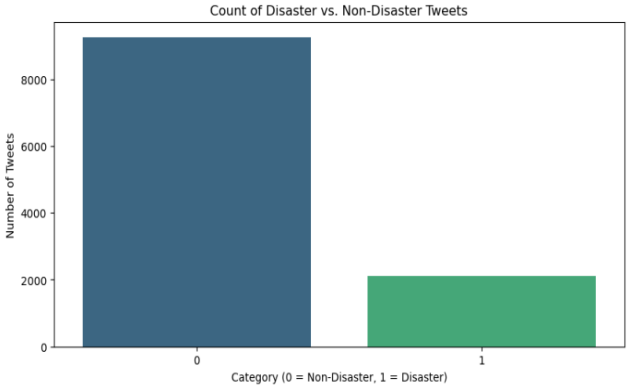
In this section, summarize the key findings of the analysis.

Keyword Analysis: The notebook identifies the most common words in disaster-related tweets.

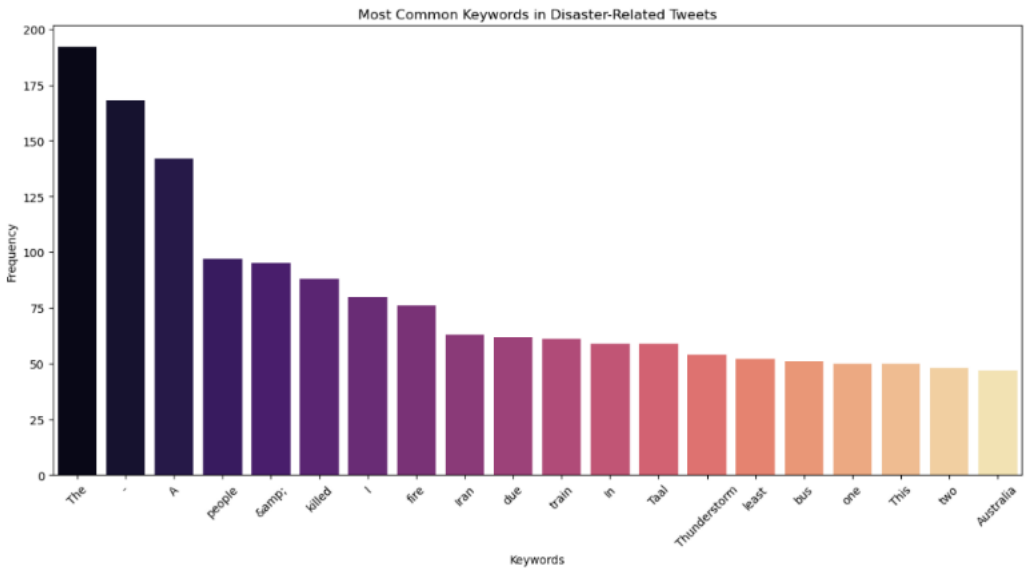
Word Frequency Analysis:

One of the main outcomes of this analysis is identifying the most frequent words in tweets labeled as disaster-related. By calculating the word frequency, the notebook highlights the most commonly used terms in these tweets.





Textual Data Analysis



Conclusion

In this analysis, we explored the critical role that Twitter plays in providing real-time information during disasters. By examining a dataset of tweets labeled as disaster-related, we identified key terms that frequently appear in such communications, including “flood,” “fire,” “earthquake,” and “help.” The visual representation of these keywords in bar plots highlights the language patterns used by individuals during emergencies. Understanding these common terms is essential for developing a machine learning model that can effectively classify tweets, allowing emergency responders to focus on relevant information amidst the overwhelming volume of social media activity during crises.

The insights gained from this project set a strong foundation for future work. The next steps involve using the identified keywords and patterns to train a machine learning model capable of accurately distinguishing between disaster-related and non-disaster tweets. By improving our ability to automate this classification, we can help emergency response teams quickly identify critical updates and needs during disasters. Ultimately, this project aims to enhance crisis management efforts, ensuring that vital information is swiftly accessed and acted upon in times of need, which can significantly improve the effectiveness of disaster response strategies.

Future Objectives for the Next Two Weeks

In the next two weeks, the primary objective will be to enhance the text preprocessing and analysis stages to ready the dataset for the development of a strong machine learning model. Key tasks will involve tokenization, lemmatization, and stemming, which aim to standardize the text by splitting it into individual words, reducing them to their base forms, and removing variations. Following this, exploratory data analysis (EDA) will be performed to identify significant patterns, relationships, and insights within the data. Additionally, feature engineering will take place to develop relevant features that could improve the model’s performance. Together, these steps will establish a solid foundation for effective model training and accurate classification of disaster-related tweets.

References

1. Gulati, N., Agarwal, A., Aggarwal, A., Bhutani, N., & Kapur, R. (2023). Ensembled multi-detector aggregation for disaster detection (EMAD). In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 593–596). IEEE.
2. Lamsal, R., & Kumar, T.V. (2023). Twitter-based disaster response using recurrent nets. In *Research Anthology on Managing Crisis and Risk Communications* (pp. 613–632). IGI Global.
3. Kanimozhi, T., Belina, V.J., & Sara, S. (2023). Classification of tweets on disaster management using random forest. In *Rajagopal, S., Faruki, P., & Popat, K. (eds.) Advances in Smart Communication and Informatics Systems (ASCIS) 2022*, Springer.
4. Rathod, J., Rathod, G., Upadhyay, P., & Vakhare, P. (2022). Disaster tweet classification using machine learning. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 523–527). IEEE.