Infosys Internship 5.0

Disaster Tweet Analyzer: Natural Language Processing for Crisis Communication

Introduction

In recent years, social media platforms, particularly Twitter, have become critical tools for disseminating information during disasters, allowing individuals to share experiences and seek assistance [1]. The vast amount of user-generated content on Twitter can be analysed to understand the nature of disasters, public sentiment, and response efforts, which can potentially save lives by facilitating timely communication [2]. This project leverages Natural Language Processing (NLP) techniques to classify disaster-related tweets, employing a robust pipeline that includes text preprocessing and feature extraction and more [3]. By utilising various classification algorithms and deep learning models , we aim to identify the most effective methods for accurately categorising tweets as “Related to Disaster” or “Not Related to Disaster.” Additionally, the role of Twitter in emergency management communication is highlighted, demonstrating its effectiveness in conveying critical information and guiding public responses during crises [4]. This approach not only enhances situational awareness but also fosters confidence in emergency management institutions.

Dataset and Methodology

Dataset

In this project, two datasets were chosen to analyse disaster-related tweets. The primary dataset, Dataset 1: Disaster Tweets, is in the form of a CSV file containing over 11,000 tweets. This dataset, available at Disaster Tweets: Real or Not?, is characterised by its lack of preprocessing and cleaning, presenting a unique challenge for analysis. Additionally, it features an imbalanced distribution, with approximately 9,000 tweets classified as non-disaster tweets and only 2,000 classified as disaster tweets. The second dataset, Dataset 2: Natural Language Processing with Disaster Tweets, consists of two CSV files: train.csv and test.csv, which have undergone preprocessing, making them more suitable for analysis and model training. This dataset is available at NLP with Disaster Tweets. Therefore, Dataset 1 will be used for learning, exploring, preprocessing, and cleaning the data, while Dataset 2 is expected to be utilised in the project going forward.

Methodology

The methodology for the Disaster Twitter Analyzer project is structured into distinct phases, encompassing data collection, preprocessing, feature engineering, model selection, and evaluation. The following outlines the steps that have been completed in the past two weeks, as well as the objectives planned for the future.

Completed Steps (Past Two Weeks) :

- Data Collection : Dataset Acquisition : Two datasets were identified and acquired for analysis, namely Dataset 1 (raw tweets) and Dataset 2 (preprocessed tweets).
- Data Exploration : Conducted exploratory data analysis (EDA) to understand the structure, characteristics, and distribution of the datasets. Insights were gained regarding the imbalanced nature of Dataset 1, which contains approximately 9,000 non-disaster tweets and 2,000 disaster tweets.
- Data Cleaning : Implemented data cleaning techniques to address missing values, remove duplicates, and filter irrelevant information from Dataset 1. This step ensured that the dataset is ready for effective analysis.
- Feature Engineering : Initial feature engineering was performed to create relevant features that could enhance model performance. This included extracting features such as tweet length, presence of hashtags, and user mentions.

Further Steps (Next Two Weeks) :

- Further Feature Engineering : Explore additional feature engineering techniques to create more complex features, such as sentiment analysis scores, temporal features based on tweet timestamps, and contextual embeddings from pre-trained language models.
- Twitter Scraping for Raw Data : Experiment with Twitter scraping methods using Twitter's API to collect more real-time tweet data. This will enhance our dataset and allow for the analysis of current trends in disaster-related tweets.
- Industry Discussion on ML and DL Techniques : Conduct a thorough review of machine learning and deep learning techniques utilised in the industry for disaster-related tweet classification. This will include examining models such as logistic regression, support vector machines, and neural networks.
- Comparison of Techniques : Perform a comparative analysis of the identified ML and DL techniques to evaluate their effectiveness, scalability, and suitability for the project. This will inform our decisions on which models to implement for tweet classification.
- Implementation of Selected Models : Based on the comparative analysis, implement and train the selected machine learning and deep learning models using the cleaned and engineered features.
- Model Evaluation : Evaluate the performance of the implemented models using metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness in classifying disaster-related tweets.

Results

The notebook explores two datasets of disaster-related tweets. Both datasets contain information about the tweet's text, location, keywords, and whether it's related to a real disaster.

Data Exploration: The analysis revealed that:

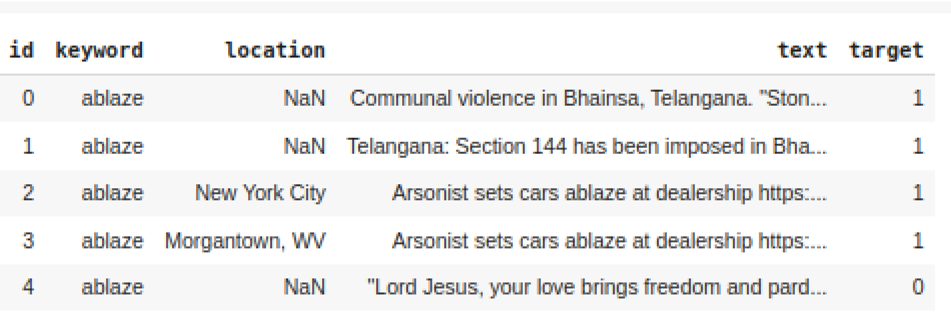
- Dataset 1 has ~7600 tweets, while Dataset 2 has ~7000.
- Both datasets have class imbalance, with more non-disaster tweets.
- The top locations in Dataset 1 were USA, UK, and Mumbai, while in Dataset 2, they were USA, New York, and London.
- The distribution of text length differs slightly between the datasets.
- The most frequent keywords were 'fire', 'flood', 'storm', etc.

Data Preprocessing: Data cleaning steps were applied to both datasets:

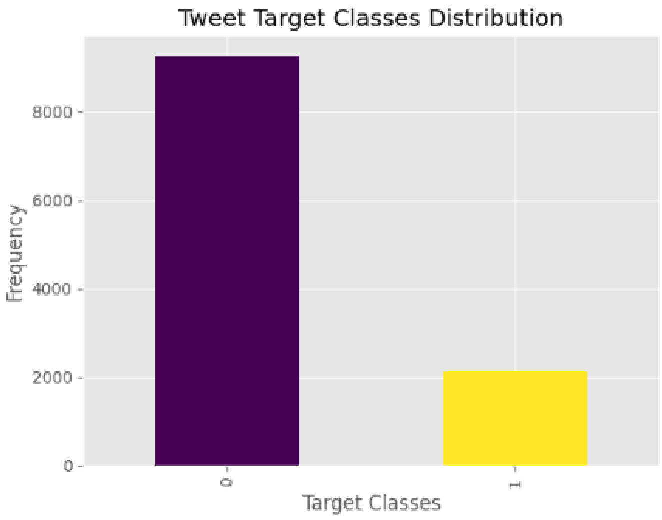
- Removal of URLs, HTML tags, and special characters.
- Decontraction and stemming of words.
- Removal of stop words and numbers.
- The 'preprocess_data' column stores the result.
- Tokenization was performed to split the cleaned text into words.

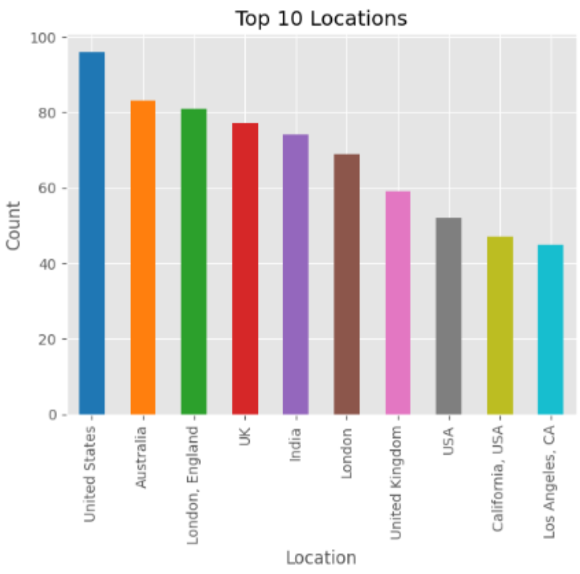
These steps are crucial for improving the performance of NLP models used for tasks like identifying disaster-related tweets.

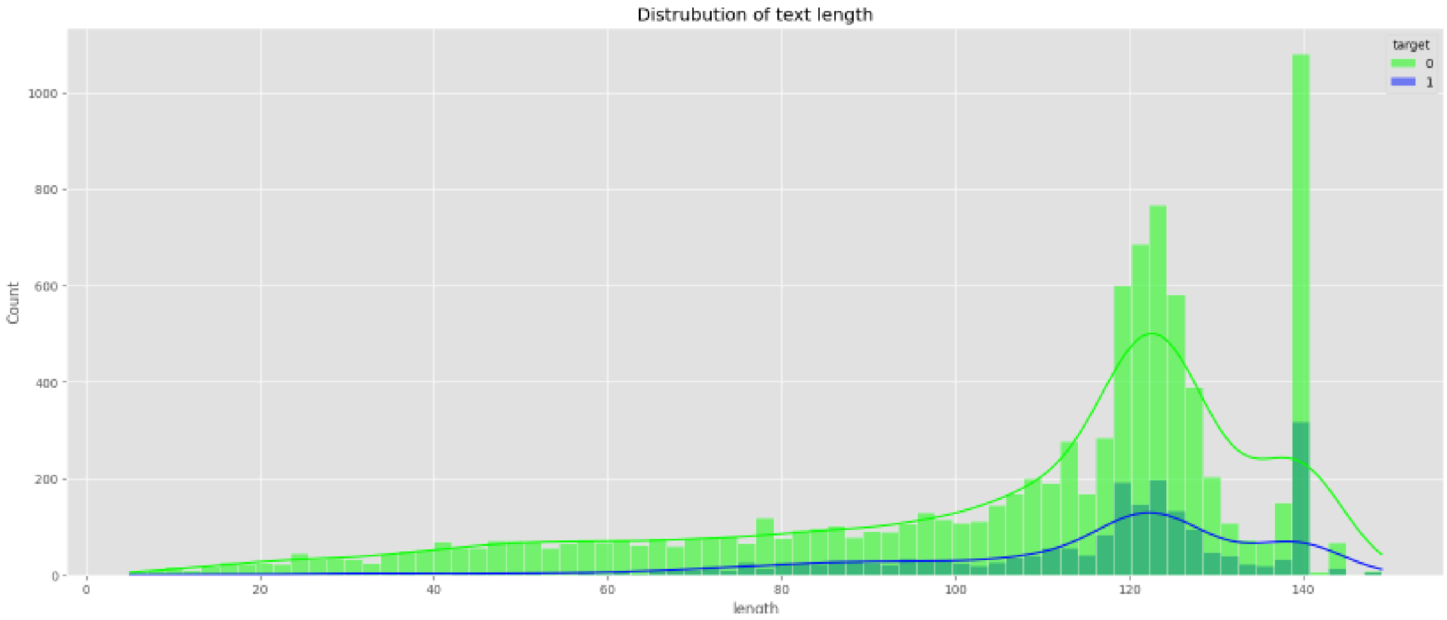
Dataset- 1



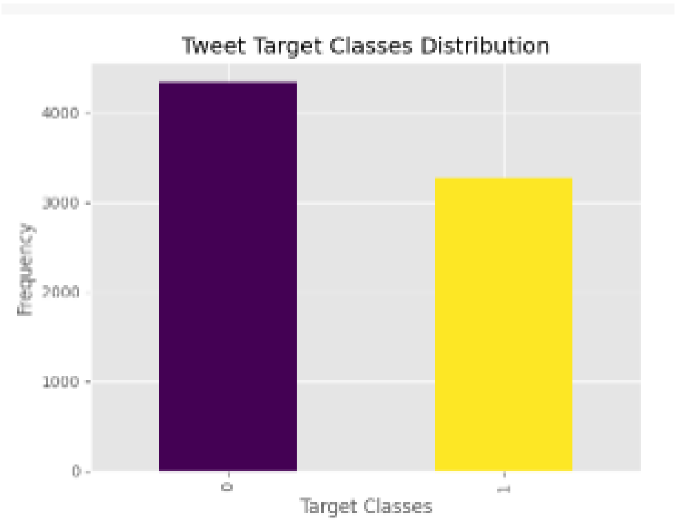
id	keyword	location	text	target
0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1
1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1
2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1
3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1
4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0

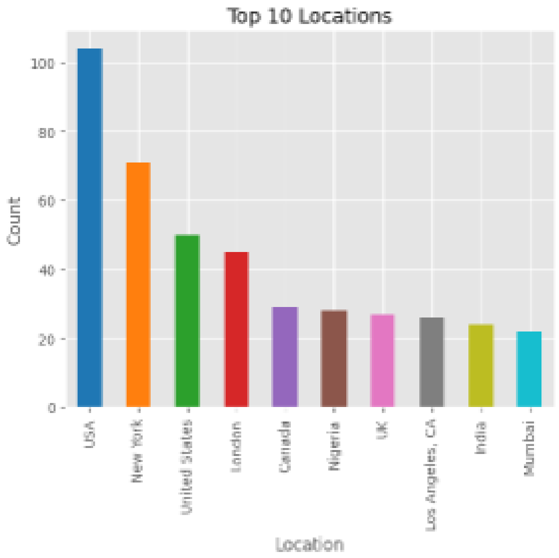


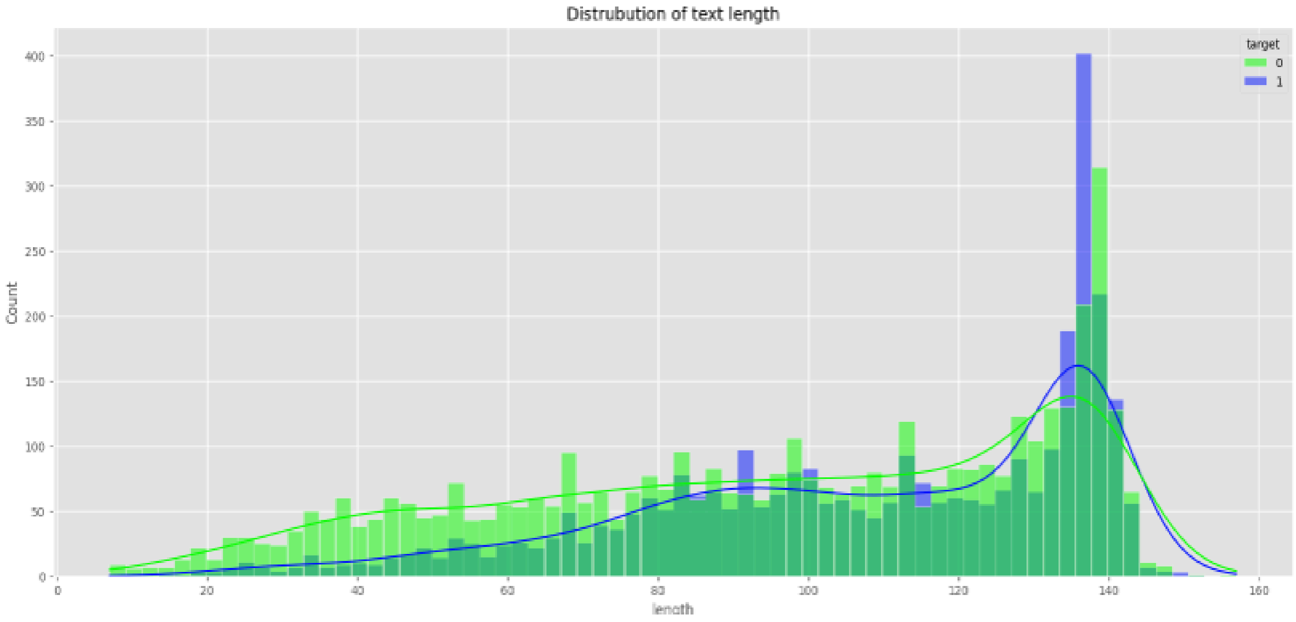



Distrubution of text length
target
0
1
Count
0
200
400
600
800
1000
0
20
40
60
80
100
120
140
length

Dataset - 2


Tweet Target Classes Distribution
Frequency
0
1000
2000
3000
4000
0
1
Target Classes


Top 10 Locations
Count
0
20
40
60
80
100
USA
New York
United States
London
Canada
Nigeria
UK
Los Angeles, CA
India
Mumbai
Location


Distrubution of text length
target
0
1
Count
0
50
100
150
200
250
300
350
400
0
20
40
60
80
100
120
140
160
length

Conclusion

In conclusion, the Disaster Twitter Analyzer project has laid a solid foundation for utilising Natural Language Processing (NLP) and Artificial Intelligence (AI) in the classification of disaster-related tweets. So far, we have successfully completed essential steps in data exploration, cleaning, and feature engineering, which are crucial for developing an effective classification model. By addressing the challenges posed by the unbalanced nature of the dataset and transforming raw tweet data into meaningful features, we have enhanced the dataset's quality for subsequent analysis. As we move forward, the insights gained during this phase will guide the implementation of machine learning and deep learning techniques, enabling real-time tweet classification. Ultimately, this system aims to provide timely and actionable information that can support authorities in responding swiftly to emerging disasters, ultimately enhancing community safety and resilience.

Future Objectives for the Next Two Weeks

In the coming two weeks, we aim to build upon the groundwork established through data exploration, cleaning, and initial feature engineering. Our objectives will focus on the following key areas:

- Further Feature Engineering : We will delve deeper into feature engineering to identify and create additional relevant features that can enhance the performance of our classification model. This may include exploring text-based features, sentiment analysis, and temporal features related to tweet timestamps.
- Twitter Scraping for Raw Data : We will experiment with Twitter scraping techniques to gather more raw tweet data, expanding our dataset. This will involve using Twitter's API to collect real-time tweets, enabling us to analyse current trends and patterns in disaster-related conversations on the platform.
- Industry Discussion on ML and DL Techniques : A thorough investigation of machine learning (ML) and deep learning (DL) techniques currently employed in the industry for similar projects will be conducted. We will analyse various methods, such as logistic regression, support vector machines, and neural networks, to understand their strengths and weaknesses in the context of disaster tweet classification.
- Comparison of Techniques : We will compile a comparative analysis of the identified ML and DL techniques, evaluating their effectiveness, scalability, and applicability to our project goals. This will help us make informed decisions regarding the models we choose to implement in our classification system.

By focusing on these objectives, we aim to enhance our understanding of the data and the methodologies available, paving the way for the successful development of a robust disaster tweet classification system.

References

[1] Wijeratne, Sanjaya & Sheth, Amit & Bhatt, Shreyansh & Balasuriya, Lakshika & Al-Olimat, Hussein & Gaur, Manas & Yazdavar, Amir & Thirunarayan, Krishnaprasad. (2017). Feature Engineering for Twitter-based Applications. 10.1201/9781315181080-14.

[2] Sharma, Anshul & Thakur, Khushal & Kapoor, Divneet & Singh, Kiran Jot & Saroch, Tarun & Kumar, Raj.(2023). Disaster Analysis Through Tweets 10.1007/978-981-19-9225-4_40.

[3] S Deepa Lakshmi1 and T Velmurugan (2023). Classification of Disaster Tweets Using Natural Language Processing Pipeline .

[4] Panagiotopoulos, Panos & Barnett, Julie & Ziaee Bigdeli, Ali & Sams, Steven. (2016). Social media in emergency management: Twitter as a tool for communicating risks to the public. Technological Forecasting and Social Change. 111. 10.1016/j.techfore.2016.06.010.

[5] https://keras.io/api/preprocessing/text/#text_to_word_sequence

[6] https://www.nltk.org/

[7] https://machinelearningmastery.com/prepare-text-data-deep-learning-keras/

[8] https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4

[9] https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/