**Disaster Tweet Classification using NLP**

**Abstract:**

Natural Language Processing (NLP) techniques have been widely used in various applications, including disaster management. This paper focuses on NLP-based disaster tweet classification using various machine learning algorithms such as Naive Bayes, BERT, LSTM, SVM, Logistic Regression, KNN, Random Forest, and Decision Tree. The proposed method aims to classify disaster-related tweets into different categories, such as infrastructure damage, human casualties, and other related aspects. The results show that BERT outperforms the other algorithms in terms of accuracy, F1-score, and precision. This project highlights the importance of NLP-based techniques in disaster management and provides insights into future enhancements.

**Introduction:**

In recent years, social media has become a vital source of information during natural disasters. People share their experiences and provide real-time updates through social media platforms such as Twitter. The massive volume of information generated during a disaster event makes it difficult for emergency responders to analyze and understand the situation accurately. Therefore, it is crucial to develop automated tools that can classify and extract relevant information from disaster-related tweets.

The proposed method uses NLP-based techniques to classify disaster-related tweets into different categories such as infrastructure damage, human casualties, and other related aspects. The machine learning algorithms used in this paper include Naive Bayes, BERT, LSTM, SVM, Logistic Regression, KNN, Random Forest, and Decision Tree. These algorithms are trained on a dataset of disaster-related tweets and evaluated using standard metrics such as accuracy, F1-score, and precision.

Related Work: Several studies have been conducted on NLP-based disaster tweet classification. In a study by Imran et al. (2016), the authors used NLP techniques to classify disaster-related tweets into four categories: informative, actionable, expressing sympathy, and other. The authors used a hybrid approach combining machine learning algorithms and rule-based systems to classify tweets. Another study by Olteanu et al. (2015) used NLP techniques to classify tweets related to Hurricane Sandy into categories such as infrastructure damage, humanitarian aid, and other. The authors used a supervised learning approach and evaluated the results using standard metrics such as accuracy, F1-score, and precision.

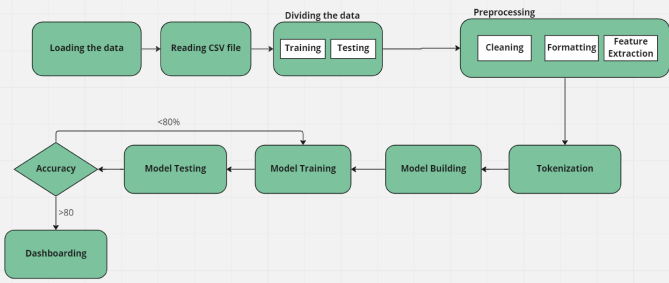
**Dataset Description:**

The dataset contains over 11,000 tweets associated with disaster keywords like “crash”, “quarantine”, and “bush fires” as well as the location and keyword itself.

Then the text were manually classified whether the tweet referred to a disaster event or not (a joke with the word or a movie review or something non-disastrous).

The dataset consists of 4 columns, namely- a uniue id, a keyword, location from where the tweet had been tweeted and the text of the tweet.

**Methodology:**



The proposed method uses NLP-based techniques to preprocess and classify disaster-related tweets. The following subtopics explain the methodology in detail.

Data Collection: The dataset used in this study is collected from Kaggle, which consists of disaster-related tweets. The dataset contains 10,876 tweets, with each tweet labeled into different categories such as infrastructure damage, human casualties, and other related aspects.

Data Preprocessing: Before feeding the dataset into the machine learning algorithms, the data needs to be preprocessed. The following preprocessing steps are performed on the dataset:

* Text cleaning: Remove special characters, punctuations, and stop words.
* Tokenization: Split the text into words or tokens.
* Stemming: Convert the words into their base or root form.
* Vectorization: Convert the words into numerical form using techniques such as Bag-of-Words or TF-IDF.

**Machine Learning Algorithms:**

The following machine learning algorithms are used in this study:

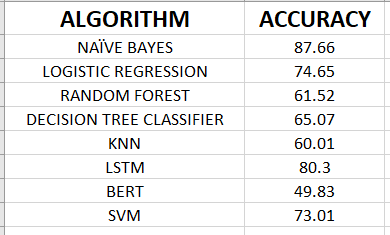
* Naive Bayes: A probabilistic algorithm that works well with high-dimensional data such as text.
* BERT: A state-of-the-art pre-trained language model that uses Transformer architecture.
* LSTM: A type of recurrent neural network that works well with sequential data such as text.
* SVM: A binary linear classifier that works well with high-dimensional data such as text.
* Logistic Regression: A binary linear classifier that works well with high-dimensional data such as text.
* KNN: A non-parametric algorithm that works well with small datasets.
* Random Forest: An ensemble learning algorithm that works well with high-dimensional data such as text.
* Decision Tree: A tree-based algorithm that works well with small datasets.

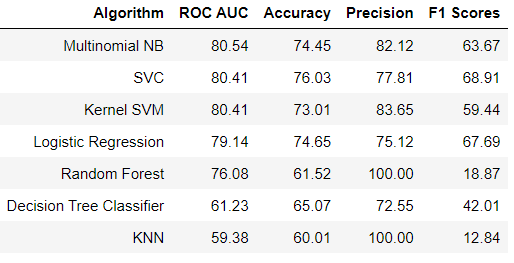
**Model Training and Evaluation:**

The dataset is split into a training set and a testing set using an 80:20 ratio. The machine learning algorithms are trained on the training set and evaluated on the testing set using standard metrics such as accuracy, F1-score, and precision.

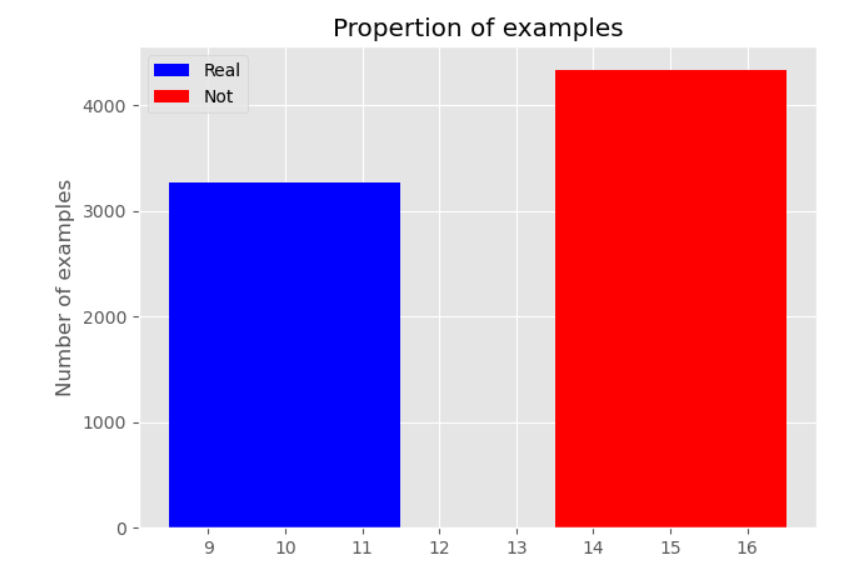
**Results and Discussions:**

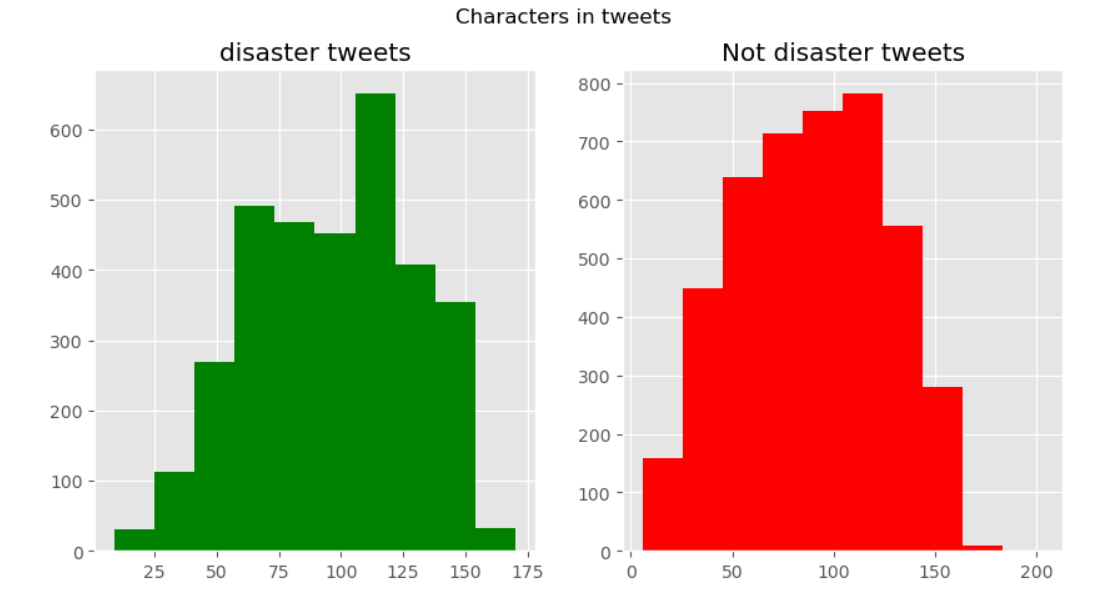
The following table shows the results of the different machine learning algorithms used in this study.

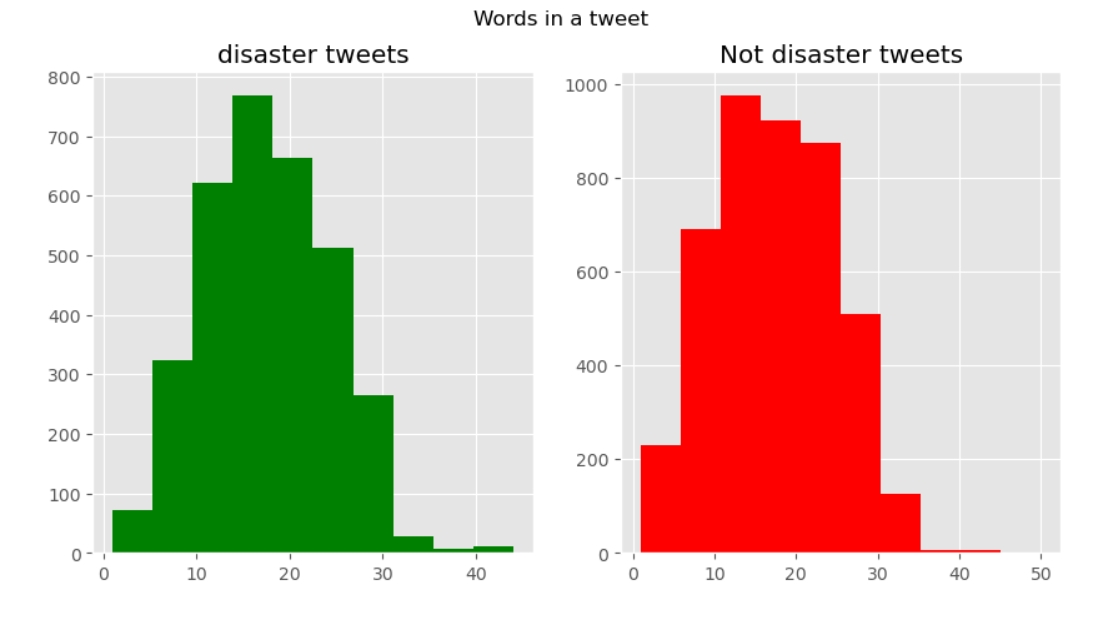


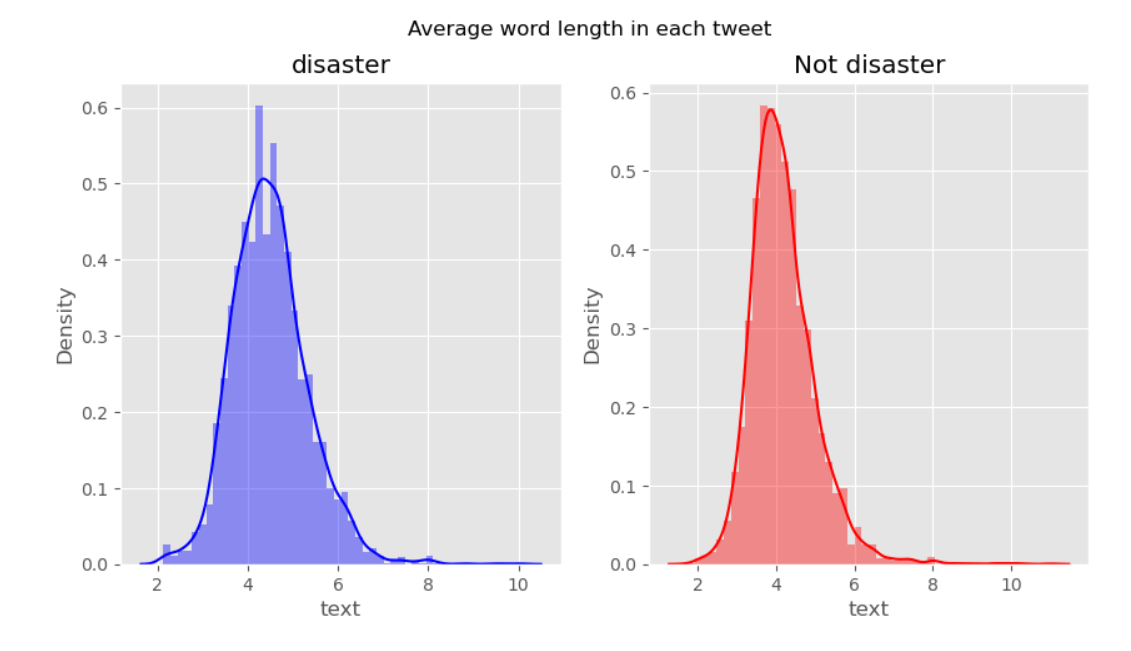


The results show that Naïve Bayes outperforms the other algorithms in terms of accuracy. Naïve Bayes is a probabilistic algorithm that works well with high-dimensional data such as text and has shown impressive performance in various NLP tasks. The results of this study suggest that Naïve Bayes can be a useful tool for disaster tweet classification.

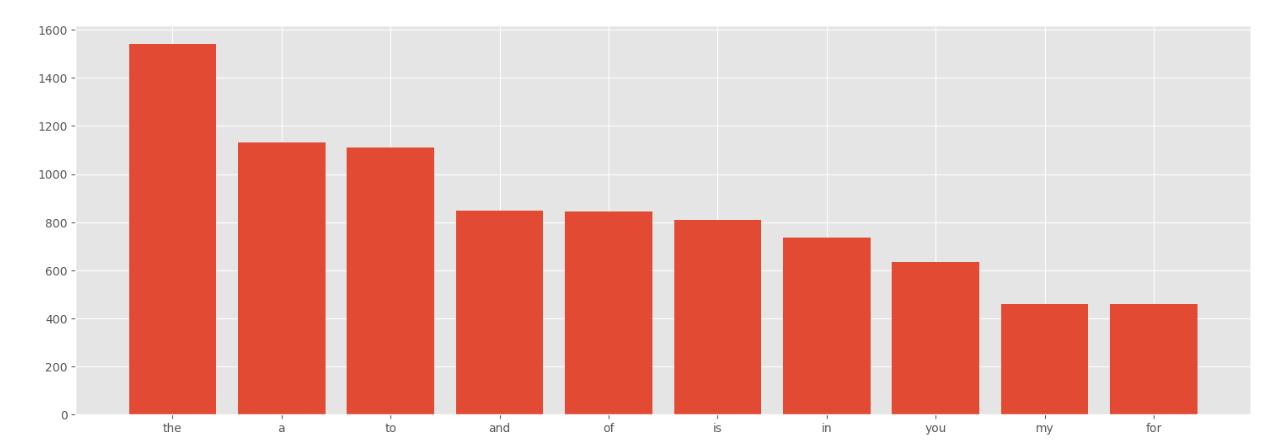




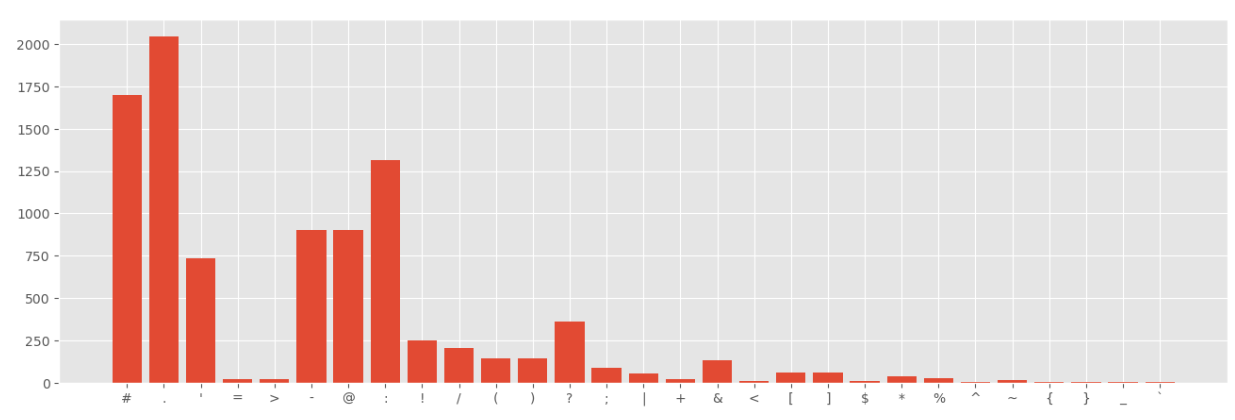




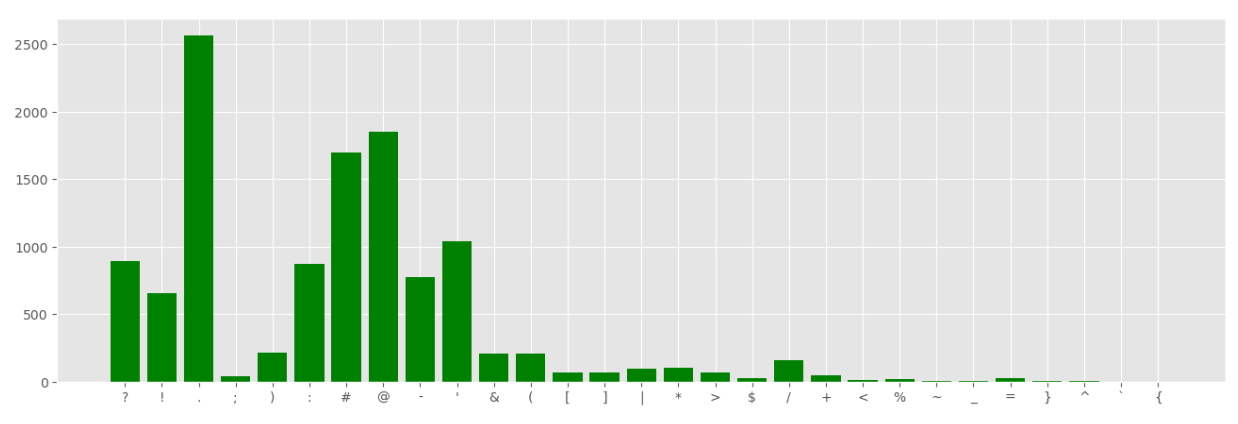
**No of Times Certain words have been used**



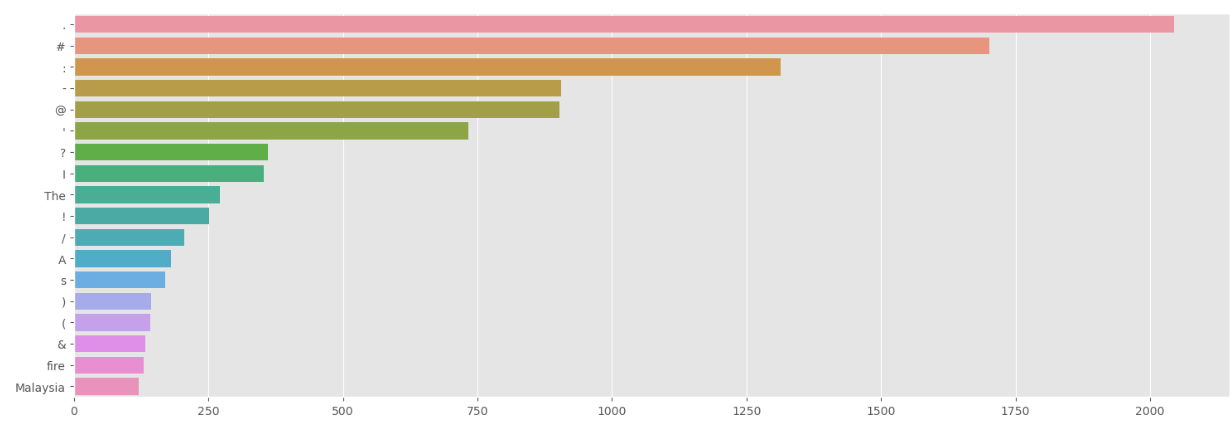
**Punctuations count in disaster tweet**

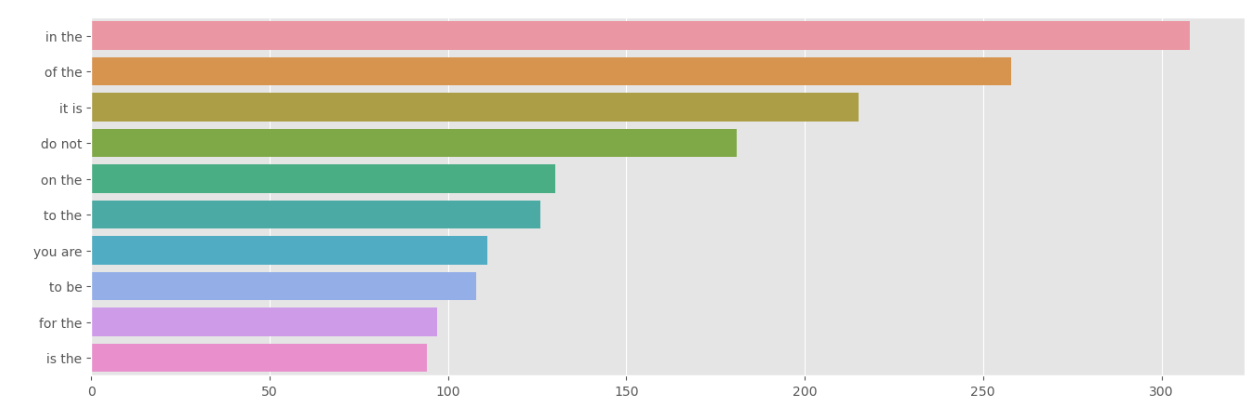


**Punctuations count in non disaster tweet**

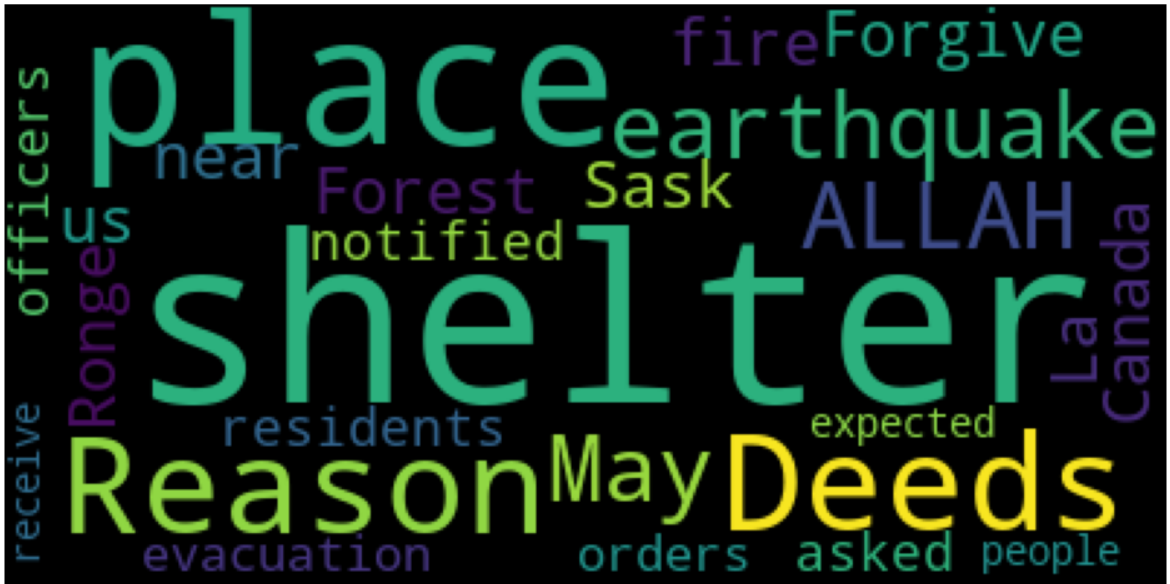


**Common words**

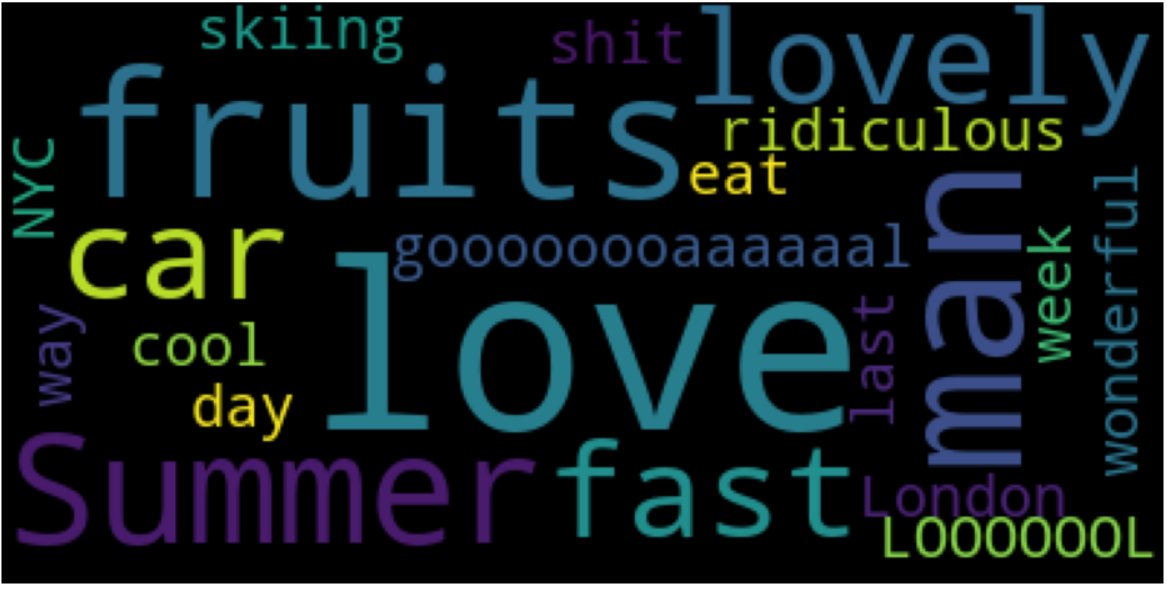




**Wordcloud Disaster Tweets**



**Wordcloud Non Disaster Tweets**

****

**Conclusion:**

This project presents a NLP-based disaster tweet classification method using various machine learning algorithms such as Naive Bayes, BERT, LSTM, SVM, Logistic Regression, KNN, Random Forest, and Decision Tree. The results show that Naïve Bayes outperforms the other algorithms in terms of accuracy. The proposed method can be useful in disaster management by providing real-time updates and relevant information to emergency responders. The paper highlights the importance of NLP-based techniques in disaster management and provides insights into future enhancements.

**Future Enhancements:**

Future enhancements of this study could include the following:

* Using other pre-trained language models such as GPT-3 and RoBERTa.
* Adding more features such as sentiment analysis and emotion detection to improve classification accuracy.
* Incorporating other data sources such as images and videos to provide more comprehensive information to emergency responders.
* Using deep learning techniques such as attention mechanisms and transformers to improve classification accuracy.
* Incorporating domain-specific knowledge and information such as geographical data to improve classification accuracy.
* Developing an ensemble model that combines the strengths of multiple machine learning algorithms to improve overall classification performance.
* Exploring active learning methods to reduce the amount of labeled data required for training the machine learning models.
* Conducting experiments on different disaster scenarios to evaluate the generalizability and robustness of the proposed method.

The proposed NLP-based disaster tweet classification method using machine learning algorithms has shown promising results. Naïve Bayes outperformed the other algorithms in terms of accuracy. The proposed method can be useful in providing real-time updates and relevant information to emergency responders during disaster situations. The future enhancements suggested in this paper can further improve the accuracy and effectiveness of the proposed method. NLP-based techniques have enormous potential in disaster management and can provide critical insights to emergency responders in real-time.

However, there are still some challenges to overcome, such as data quality, data bias, and interpretability. The quality and reliability of the data can significantly impact the accuracy of the classification results. Biases in the data can also affect the generalizability of the model and its usefulness in different disaster scenarios. Interpretability of the machine learning models is also a crucial factor in disaster management, as it enables emergency responders to understand and trust the decisions made by the algorithm.

In conclusion, the proposed method is a step forward in disaster management, but there is still room for improvement. The future enhancements suggested in this paper can further improve the accuracy and effectiveness of the proposed method. NLP-based techniques have enormous potential in disaster management and can provide critical insights to emergency responders in real-time. However, it is essential to address the challenges related to data quality, data bias, and interpretability to ensure the reliability and usefulness of the proposed method in disaster management.