REVIEW ANALYSIS OF DISASTER TWEETS BASED ON SOCIAL MEDIA

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF
REQUIREMENTS FOR THE AWARD OF THE DEGREE

MASTER OF SCIENCE IN DATA ANALYTICS

OF

MAHATMA GANDHI UNIVERSITY, KOTTAYAM

BY

FRANCE ANTONY JOSEPH

220011023642

ESTD 1964

Department of Mathematics & Statistics

BASELIUS COLLEGE KOTTAYAM

(NAAC Re-accredited with A++ Grade)

Kottayam, Kerala–686001

2024

A Project Report on

REVIEW ANALYSIS OF DISASTER TWEETS BASED ON SOCIAL MEDIA

SUBMITTED IN PARTIAL FULFILMENT OF REQUIREMENT
FOR THE AWARD OF THE DEGREE

MASTER OF SCIENCE IN DATA ANALYTICS

OF

MAHATMA GANDHI UNIVERSITY, KOTTAYAM

By

FRANCE ANTONY JOSEPH

220011023642

Under the guidance ands upervision of

Ms. Ross Cyriac
Lecturer
Department of Mathematics & Statistics
Baselius College Kottayam

ESTD 1964

BASELIUS COLLEGE KOTTAYAM

(NAAC Re-accredited with A++ Grade)

Kottayam Kerala–686001

2024

BASELIUS COLLEGE KOTTAYAM

(NAAC Re-accredited with A++ Grade)

Kottayam, Kerala – 686 001

ESTD 1964

Department of Mathematics & Statistics

CERTIFICATE

This is to certify that the project work entitled **“REVIEW ANALYSIS OF DISASTER TWEETS BASED ON SOCIAL MEDIA”** is a bonafide record of work done by **FRANCE ANTONY JOSEPH, 220011023642,** in partial fulfilment of the requirements for the award of Degree of **MASTER OF SCIENCE IN DATA ANALYTICS** during the academic year 2023-2024.

Ms. Ross Cyriac
Lecturer
Dept of Mathematics & Statistics
Baselius College Kottayam

Ms. Preetha Mathew
Head Of The Department
Dept of Mathematics & Statistics
Baselius College Kottayam

External Examiner
Place: Kottayam
Date:

DECLARATION

I, **FRANCE ANTONY JOSEPH (RegNo:220011023642)** hereby declare that this project work entitled **“REVIEW ANALYSIS OF DISASTER TWEETS BASED ON SOCIAL MEDIA”** is a record of original work done by me under the guidance of **Ms. Ross Cyriac**, Lecturer, Department of Mathematics and Statistics and the work has not formed the basis for the award ofanydegree or diploma or similar title to any candidate of any university subject.

Place: Kottayam **FRANCE ANTONY JOSEPH**

Date:

ACKNOWLEDGEMENT

This project is not complete if one fails to acknowledge all who have been instrumental in the successful completion of the project. If words were to be the symbol of undiluted feelings and token of gratitude then let the words play the heralding role of expressing my gratitude.

First of all, I thank the **“God Almighty”** for his immense grace and blessings in my life and at each stage of this project.

I express my sincere and profound gratitude to **Prof. Dr. Biju Thomas**, Principal, Baselius college Kottayam, for providing all the facilities during the period of the project.

I extend my gratitude to **Ms. Preetha Mathew**, Head of the Department, Department of Mathematics and Statistics, who is a constant source of inspiration and whose advice helped me to complete this project successfully.

I express my deep sense of gratitude to my internal project guide **Ms. Ross Cyriac**, Lecturer, Department of Mathematics and Statistics, for her profound guidance for the successful completion of this project.

With great enthusiasm I express my gratitude to all the faculty members of the Department of Mathematics and Statistics for their timely help and support.

I would like to thank **SMEClabs, Kaloor**, for providing me with an invaluable learning experience during my project in Data Science. It has been a pleasure to work with the staff, and I have gained much practical knowledge about Data Science.

Finally, I express my deep appreciation to all my friends and family members for the moral support and encouragement they have given to complete this project successfully.

FRANCE ANTONY JOSEPH

ABSTRACT

Twitter gaining more popularity among several other social media sites, and generating massive amount of data generated by social media present a unique opportunity for disaster analysis. Generates millions of massive Tweets each day and its real-time characteristics, employing sentiment analysis can bring more insights on the impact of tweets. The objective of the project involves whether a tweet is an disaster tweet or not using the content shared on various social media platforms during and after disaster events. This process utilizes natural language processing techniques to discern the public's emotional responses to disasters, whether they be natural calamities like hurricanes and earthquakes or man-made crises.

The Social Media Disaster Tweets dataset, sourced from Data World, presents a collection of tweets pertaining to various calamitous events shared across social media platforms. With a focus on disaster-related discourse, the dataset encompasses a single column comprising 10,876 rows of textual data. Each tweet offers insights into the sentiments, reactions, and details surrounding a wide spectrum of disasters, including natural phenomena like hurricanes and earthquakes, as well as human-made crises such as accidents and conflicts. Serving as a valuable resource for research and analysis, the dataset enables the training of algorithms to discern and categorize different types of disaster-related tweets.

This project involves natural language processing (NLP) and Deep Learning based Artificial Neural Network (ANN). Deep learning based advanced Ai algorithms for text classification. Preprocessing and filtering of text is done with Natural language toolkit (NLTK) and Pandas.

The programming environment utilized for this project includes Google Collab and Visual Studio Code, with a minimum Python version 3.9.0.

TABLE OF CONTENTS

1. INTRODUCTION

1. INTRODUCTION

In today's interconnected world, social media platforms serve as vital channels for communication during times of disaster. Whether it's natural calamities like earthquakes, floods, or human-made emergencies such as accidents or conflicts, people turn to social media to share information, seek help, and provide support. Understanding the dynamics of how information flows through these platforms during disasters is crucial for emergency responders, policymakers, and researchers alike. This project aims to delve into the realm of disaster tweets on social media, exploring various aspects such as the types of information shared, the spread of misinformation, the role of social networks in disseminating critical updates, and the impact of these communications on disaster response and recovery efforts.

Disaster tweets on social media play a crucial role in communication during crises, providing real-time updates, emotional support, and opportunities for community engagement. However, they also present challenges in terms of verifying information and managing the spread of misinformation. By studying these dynamics, researchers and emergency responders can work towards harnessing the power of social media for more effective disaster management and response.

1.1. OVERVIEW OF THE PROJECT

Social media platforms have become indispensable tools for communication during times of crisis. From natural disasters to human-made emergencies, people turn to platforms like Twitter to share information, seek help, and express their experiences and emotions. The data set is from Data World. The dataset consists of a nearly 10876 text input data. The dataset of disaster tweets on social media provides an invaluable resource for developing and training NLP models for sentiment analysis. These models can play a crucial role in understanding public sentiment during disasters, enhancing situational awareness, and facilitating more effective communication and response strategies.

Building classification models using various algorithms such as Random Forest, Logistic Regression, Support Vector Machine (SVM), and Artificial Neural Network (ANN) on the dataset of disaster tweets can help determine which algorithm performs the best in terms of accuracy.

An Artificial Neural Network (ANN) is a computational model inspired by the structure and functioning of biological neural networks found in the human brain. These networks are fundamental to machine learning and deep learning, enabling tasks such as pattern recognition, classification, regression, and decision-making. Typically, an ANN consists of an input layer, one or more hidden layers, and an output layer. The input layer receives the input data, with each part of this layer conveying specific details about the input. Hidden layers, situated in the middle, process the input using connections with weights and specialized functions, adding complexity to the system. Each part of a hidden layer represents something the network has learned from the input. Finally, the output layer provides the final result of the network's computation. In constructing an ANN, it is crucial to choose appropriate activation functions. For instance, Rectified Linear Unit (ReLU) activation functions are often used for hidden layers, while the sigmoid function is suitable for the output layer, particularly in cases of multi-class classification.

1.2. RELEVANCE OF THE PROJECT

The relevance of a project on disaster tweets on social media lies in its profound impact on both disaster management and public safety. In simple terms, this project addresses the critical need to understand how people communicate and seek help during disaster susing platforms like Twitter, Facebook, and others. By analysing the content and sentiment of these tweets, we can gain insights into the real-time needs and concerns of those affected by disasters. Moreover, by identifying patterns in the spread of information and misinformation on social media during disasters, we can develop strategies to improve communication and combat rumours, ultimately enhancing public safety and resilience in the face of crises. In essence, this project is all about harnessing the power of social media to save lives and mitigate the impact of disasters on communities.

The analysis of social media data during natural disasters can be challenging due to the sheer volume of data generated and the need to quickly identify relevant information. Additionally, tweets are often short, informal, and contain non-standard language, making them difficult to analyse using traditional NLP techniques. As a result, there is a need for more advanced NLP techniques that can accurately classify disaster-related tweets and extract relevant information in real-time. The dataset provided for this challenge consists of a collection of tweets that have been labelled as either “disaster” or “not disaster”.

2. SYSTEM STUDY

2. SYSTEM STUDY

2.1. PROPOSED SYSTEM

The initial phase in examining the Social Media Disaster dataset includes gathering the data.This can be achieved by either scraping the review data from Kaggle or utilizing a publicly accessible dataset such as the one from Data World. Subsequently, preprocessing the data becomes crucial. This includes tasks like cleaning the data by eliminating special characters or stop words, and transforming the text into numerical representations suitable for input into an ANN model. Once the data preprocessing is completed, the next step involves building machine learning models capable of analysing the sentiment expressed in the tweets.

Following the preprocessing of the data, the Machine Learning model is trained utilizing the prepared data to classify tweets into positive and negative categories. Subsequently, the performance and accuracy of the trained models are assessed. This involves splitting the dataset into training and testing sets, training the machine learning models on the training set, and then evaluating their performance using the test set.

2.2. CHALLENGES IN MODEL BUILDING

Preprocessing the Social Media Disaster dataset is crucial prior to training a Natural Language Processing (NLP) model. This involves several tasks, including data cleaning, stopword removal, and converting textual data into a numerical format suitable for analysis. Additionally, it's important to address potential class imbalances within the dataset. This imbalance may result in a significant disparity between the number of positive and negative tweets, posing challenges for training an accurate model. Such an imbalance could lead the model to be biased towards the majority class, impacting its ability to accurately classify tweets. This can lead to poor performance when the model is applied to new, unseen data.

In Summary, creating and training NLP models for analyzing the dataset can be tough. It involves careful steps like preparing the data, adjusting settings, and considering the resources needed. It's crucial to review the outcomes carefully and understand the limitations and difficulties in building the models.

2.3. LIMITATIONSOFTHESYSTEM

The model's effectiveness relies on the size of the training dataset. The project's dataset is relatively small, which can reduce the accuracy of predictions and potentially lead to over fitting. Moreover, the dataset includes only a limited number of features, potentially excluding other features that could enhance the model's accuracy.

The model's performance is assessed using the dataset it was trained on, which might not offer an accurate estimation of how well it performs on new, unseen data. The project focuses solely on Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and Support Vector Machines (SVM) for predicting whether it is Social media disasters tweets or not, with Artificial Neural Networks (ANN)employed for performance comparison.

3. SYSTEM DESIGN

3. SYSTEM DESIGN

3.1. ABOUT THE LANGUAGE

Python is an interpreted, object-oriented, high-level programming language renowned for its dynamic semantics and ease of use, originally conceived by Guido van Rossum in 1991. With a primary focus on readability, Python employs simple syntax and clear rules, facilitating swift program development and seamless integration of disparate components. Its support for modules and packages enables efficient organization and code reuse, enhancing project maintainability. Python's versatility extends across diverse domains, from web development and data analysis to artificial intelligence and scientific computing. It boasts a vibrant community and extensive documentation, fostering collaborative learning and problem-solving. Available for free and compatible with major operating systems, Python stands as a versatile and accessible tool for programmers world wide .Analysis and model building are conducted utilizing Python version 3.9.0.

3.2. ABOUT THE EDITOR

Google Colab, also known as Google Colaboratory, stands as a versatile tool offered by Google, enabling users to write and execute Python code directly within a web browser. Widely embraced by data scientists and machine learning practitioners, its popularity stems from being both free and granting access to specialized computing resources such as Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), which significantly accelerate the training process for machine learning models. Through Colab, users can author and execute Python code within a unique document format known as a Jupyter Notebook, which seamlessly integrates text, code, and visual elements, fostering efficient data analysis and idea exploration. Furthermore, Colab integrates seamlessly with Google Drive, facilitating easy storage and sharing of work. Equipped with essential Python libraries like NumPy, pandas, TensorFlow, and scikit-learn, it streamlines the development process. Additionally, Colab leverages powerful hardware resources to expedite tasks, particularly in training complex machine learning models. Moreover, it supports real-time collaboration, enabling multiple users to collaborate effortlessly on projects, thus enhancing productivity and teamwork.

3.3. DATA SET FOR THE STUDY

The Social Media Disaster Dataset, sourced from Data World, comprises tweets related to various disasters shared by users across social media platforms. With a single column and 10,876 rows, the dataset primarily encapsulates textual information, offering insights into the sentiments, reactions, and details surrounding diverse calamitous events. Each row represents a distinct tweet, ranging from natural disasters like hurricanes and earthquakes to human-made crises such as accidents or conflicts. This dataset serves as a valuable resource for training computer programs to discern and categorize different types of disaster-related tweets, aiding in the development of algorithms capable of identifying and analyzing relevant content amidst the vast expanse of social media data.

3.4. ABOUT NLP

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language in a way that is both meaningful and contextually relevant. At its core, NLP seeks to bridge the gap between human communication and computational systems, enabling machines to comprehend and interact with natural language data. NLP algorithms employ a variety of techniques to process and analyze text, including tokenization, part-of-speech tagging, named entity recognition, syntactic parsing, and sentiment analysis, among others. These techniques enable NLP systems to perform a wide range of tasks, such as language translation, text summarization, sentiment analysis, question answering, and text generation.

NLP has numerous applications across various industries and domains. In healthcare, NLP can be used to extract insights from medical records, assist in clinical decision-making, and improve patient care. In finance, NLP enables sentiment analysis of news articles and social media posts to inform investment decisions and market trends. In customer service, chatbots and virtual assistants utilize NLP to understand and respond to user queries in natural language, enhancing the customer experience. Overall, NLP continues to play a crucial role in advancing AI capabilities and enabling more intuitive and intelligent interactions between humans and machines.

3.5. ABOUT ANN

Artificial Neural Networks (ANNs) represent a fundamental architecture within the realm of artificial intelligence (AI) and machine learning. Inspired by the structure and functioning of the human brain, ANNs consist of interconnected nodes, or neurons, organized into layers. These layers typically include an input layer, one or more hidden layers, and an output layer. Each neuron receives input signals, performs a weighted sum of these inputs, applies an activation function to compute an output, and then transmits this output to the neurons in the subsequent layer. Through a process known as forward propagation, data flows through the network, undergoing transformations at each layer to produce a final output.

Despite their effectiveness, ANNs also present certain challenges. Designing and training neural networks requires careful consideration of hyper parameters, network architecture, and training data, and can be computationally intensive. Furthermore, over fitting, where the model learns to memorize the training data rather than generalize to unseen data, is a common concern that necessitates regularization techniques and appropriate validation procedures.

Overall, ANNs represent a powerful and versatile tool in the field of machine learning, with the potential to revolutionize numerous industries and domains through their ability to learn complex patterns and relationships from data.

4. ALGORITHMS USED

4. ALGORITHMS USED

Artificial intelligence encompasses the concept of machines performing tasks intelligently by simulating human behaviours and cognitive processes. Within this domain, Machine Learning servesasa specialized area focusing on constructing AI-powered applications. Deep Learning, in turn, resides as a subset of Machine Learning, leveraging extensive datasets and intricate algorithms to train models. In the context of this dataset, Machine Learning algorithms are utilized for analysis and processing**.**

The main algorithms used in this project are:

- ✓ Random Forrest Classifier
- ✓ Logistic regression
- ✓ Support Vector Machine (SVM)
- ✓ Decision Tree Classifier
- ✓ Artificial Neural Network (ANN)

Random Forest Classifier

The Random Forest Classifier constructs a collection of decision trees by randomly selecting subsets from the training set. It then combines the predictions from these individual trees to determine the final classification of a test object. Key parameters for tuning the Random Forest Classifier include the total number of trees to generate and decision tree-specific parameters such as minimum split criteria. As its name suggests, a random forest comprises numerous decision trees that function collectively. Each tree predicts a class, and the most commonly predicted class across all trees becomes the model's overall prediction. The underlying principle driving the effectiveness of random forests is the concept of the wisdom of crowds. In essence, this means that a large ensemble of relatively uncorrelated models (trees) working together as a committee tends to out per for many single constituent model. A random forest classifier is a popular machine learning algorithm used for classification tasks. It is an ensemble learning method that combines multiple decision trees to make a prediction.

Logistic regression

Logistic regression, a technique derived from statistics and adapted for machine learning, serves as a primary method for solving binary classification problems, where there are only two possible class values. The logistic model, also known as the logit model, is employed to model the probability of a specific class or event occurrence, such as pass/fail, win/lose, alive/dead, or healthy/sick. This algorithm assigns observations to discrete classes and is commonly utilized in various classification scenarios, including determining email spam or legitimate messages etc. Logistic regression employs the logistic sigmoid function to transform its output, returning a probability value. By leveraging the concept of probability, logistic regression serves as a predictive analysis algorithm, facilitating informed decision-making in classification tasks.

Instead of utilizing a linear function, the cost function can be represented by the sigmoid function, also known as the logistic function. This logistic function is applied to the equation of a straight line. Additionally, a significant threshold is established, where values exceeding this threshold are interpreted as 1, while those below it are interpreted as 0.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful and versatile supervised learning algorithm used for classification and regression tasks. Its primary objective is to find the optimal hyperplane that best separates data points belonging to different classes in a high-dimensional space. SVM achieves this by identifying a decision boundary that maximizes the margin, which is the distance between the hyper plane and the nearest data points, known as support vectors.

In addition to classification tasks, SVMs can also be extended to handle regression tasks by minimizing the error between predicted and actual values while still maximizing the margin.SVM is particularly effective in scenarios with high-dimensional data and when the number of features exceeds the number of samples. It is also robust against over fitting, as it seeks to maximize the margin while penalizing misclassifications.

Overall, SVM is a widely used and versatile algorithm that excels in a variety of applications, including text categorization, image classification, bioinformatics, and financial forecasting, making it a valuable tool in the machine learning toolbox.

Decision Tree Classifier

The Decision Tree algorithm is a versatile and intuitive supervised learning technique used for both classification and regression tasks. It operates by recursively partitioning the input space into regions, with each partition being associated with a specific class label or a predicted value. Decision trees are constructed based on a set of hierarchical if-else conditions, where each internal node represents a decision based on a feature, and each leaf node corresponds to a class label or a predicted value.

Decision trees find applications in various domains, including medicine, finance, and marketing. For instance, in medicine, decision trees can be used to assist in disease diagnosis by identifying the most relevant symptoms for differentiating between diseases. In finance, decision trees can help in credit scoring and fraud detection by predicting the likelihood of default or fraudulent activity based on customer attributes. Overall, decision trees represent a flexible and interpretable machine learning approach that can provide valuable insights and predictions across a wide range of problems.

Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) serve as digital counterparts to brain cells, collaborating to process and interpret information. Comprising artificial neurons, known as units, organized into layers, these networks adapt to handle various tasks. Layers in a neural network typically include the input layer, hidden layer(s), and output layer. The input layer receives data from the external environment, which progresses through the hidden layers for processing and comprehension. Subsequently, the output layer generates the network's response or prediction based on the input data. Interconnections between neurons in adjacent layers are governed by weights, determining the influence of one neuron on another. As data traverses through the network, learning occurs, refining the network's understanding and culminating in the generation of an output from the output layer. Through this iterative process, neural networks exhibit the ability to learn and improve their performance over time.

5. ANALYSIS & INTERPRETATION

5. ANALYSIS AND INTERPRETATION

5.1. PURPOSE AND WORKFLOW

The project progresses through several stages, beginning with data management, followed by visualization, analysis, and classification modelling. Exploratory Data Analysis (EDA) plays a pivotal role in this process, wherein datasets are meticulously examined to discern their primary characteristics, often utilizing visual methods. EDA serves to extract insights beyond formal modelling or hypothesis testing, providing valuable inference about the underlying data. In this study, the objective is to evaluate various classification algorithms on the dataset and scrutinize the outcomes. The workflow encompasses the following steps:

- Dataset loading & EDA
- Visualization & Data analysis
- Data cleaning & processing
- Model Building
- Classification modelling

The natural language processing is done to convert the text for the model building. After that a classification model is tried to build by using different classification algorithms and those which highest accuracy is selected.

5.2. PYTHON IN DATA SCIENCE PROJECT

Python is a versatile and user-friendly programming language known for its simplicity and readability. Python's design emphasizes code clarity and conciseness, making it an ideal choice for beginners and experienced programmers alike. With a rich ecosystem of libraries and frameworks, Python is widely used for web development, data analysis, machine learning, artificial intelligence, scientific computing, and more. Its popularity stems from its ease of learning, extensive documentation, and robust community support. Overall, Python serves as a powerful tool for solving a diverse range of programming tasks efficiently and effectively. These libraries provide pre-built tools and functionalities, streamlining the data analysis process and empowering users to extract insights efficiently. Given Python's versatility and widespread adoption, it was only a matter of time before it became a go-to choice for data analysis tasks.

Modules and Packages

In the analysis and visualization of datasets, Python offers a plethora of built-in modules and packages that facilitate various tasks. Some of the commonly used modules in this context include:

- **NumPy** : A fundamental package for numerical computing in Python,
- **Pandas**: A powerful library for data manipulation and analysis, offering data structures like Data Frame and Series
- **Matplotlib**: A versatile plotting library for creating static, interactive, and animated visualizations in Python.
- **Seaborn**: Built on top of Matplotlib, Seaborn provides a high-level interface for creating attractive and informative statistical graphics.
- **Scikit-learn**: A comprehensive machine learning library in Python, offering a wide range of algorithms for classification, regression, clustering, dimensionality reduction, and more. It also provides tools for model selection, evaluation, and preprocessing.
- **wordcloud**: Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.
- **nltk**: The Natural Language Toolkit (NLTK) is a Python package for natural language processing. NLTK is one of the leading platforms for working with human language data and Python, the module NLTK is used for natural language processing.
- **TensorFlow**: It is an open-source software library developed by Google Brain which allows developers to build and train machine learning models by creating computational graphs, which represent the operations that the model performs.
- **Keras**: It is an open-source deep learning framework written in Python. It is designed to enable fast experimentation with deep neural networks, while also providing a user- friendly and modular API.

5.3. DATA CLEANING AND PREPROCESSING

Initially, the dataset is loaded using the read_csv() function from the pandas library, which enables data to be read into a Data Frame directly from a remote URL. To gain insights into the dataset's structure and content, functions and attributes provided by pandas are leveraged. The head() function, for instance, offers a preview of the dataset by displaying the first few rows, facilitating a quick overview of the data. Additionally, the shape attribute reveals the dimensions of the dataset, indicating the number of rows and columns.

The social media disaster dataset comprises a total of 10,876 observations, with each observation containing 1 feature variable or attribute. Utilizing the data.info() method in pandas, we can ascertain the number of non-null values in each column, along with the respective data types associated with each column. This information is essential for understanding the completeness of the dataset and the nature of the data stored within it, enabling further exploration and analysis to be conducted effectively.

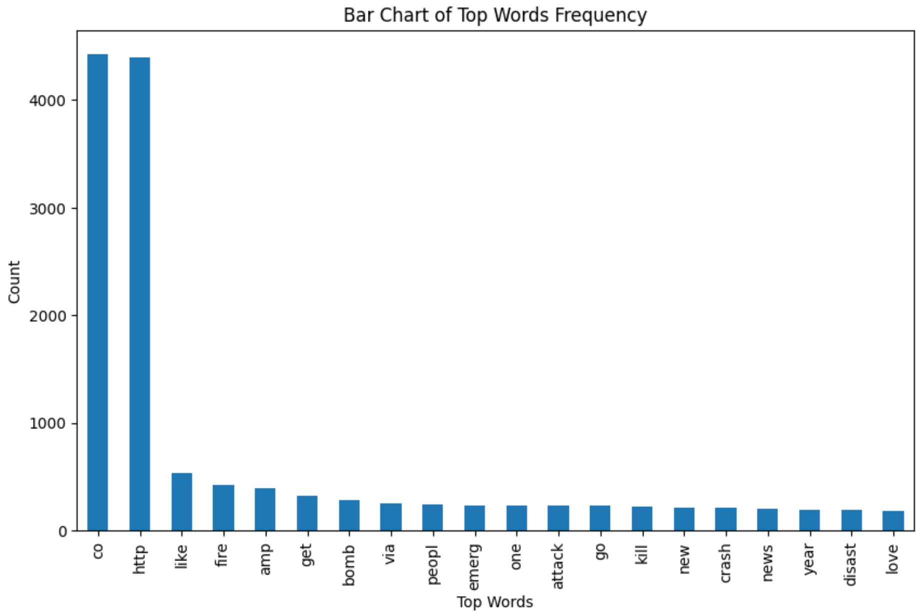
The dataset is examined for missing values by employing the is null() function within the pandas library. Upon inspection, it is determined that the dataset does not contain any missing values across any of its variables or features. This absence of missing values suggests that the data has been meticulously collected or subjected to thorough cleaning prior to its publication. Ensuring the absence of missing values is crucial for conducting a robust analysis, as missing data can potentially skew results and compromise there liability of findings. Therefore, the confirmation of a clean dataset lays a solid foundation for subsequent analysis tasks.

5.4. DATA VISUALIZATIONS & INTERPRETATIONS

Data visualization is a crucial aspect of data analysis, aiming to unveil patterns, trends, and correlations within datasets by presenting them visually. Python boasts several powerful graphing libraries equipped with diverse features, catering to various visualization needs. Whether it's crafting interactive, dynamic, or extensively customized plots, Python offers an array of excellent tools. Here are a few popular plotting libraries commonly employed in projects:

- Matplotlib
- Seaborn

- **Barchart of Top words Frequency**

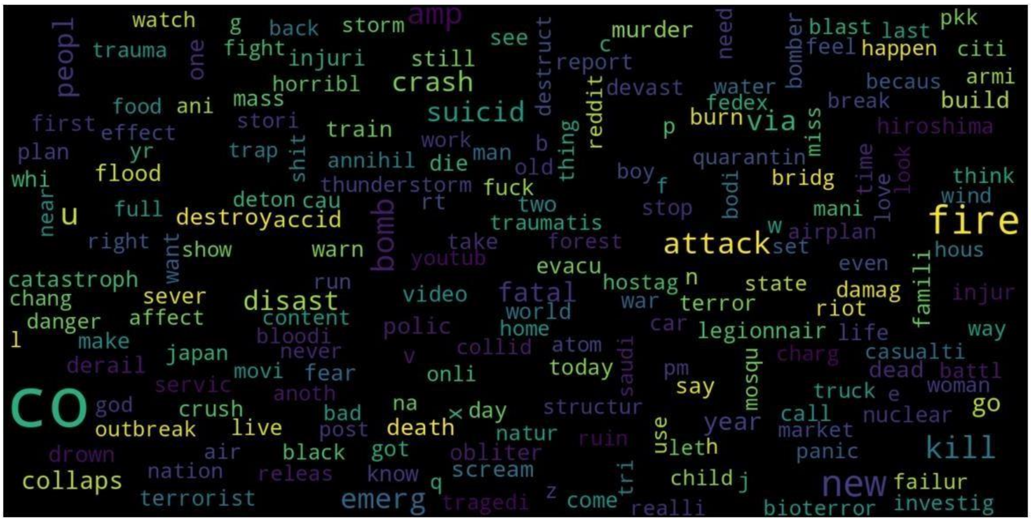


Library : matplotlib & Worldcloud

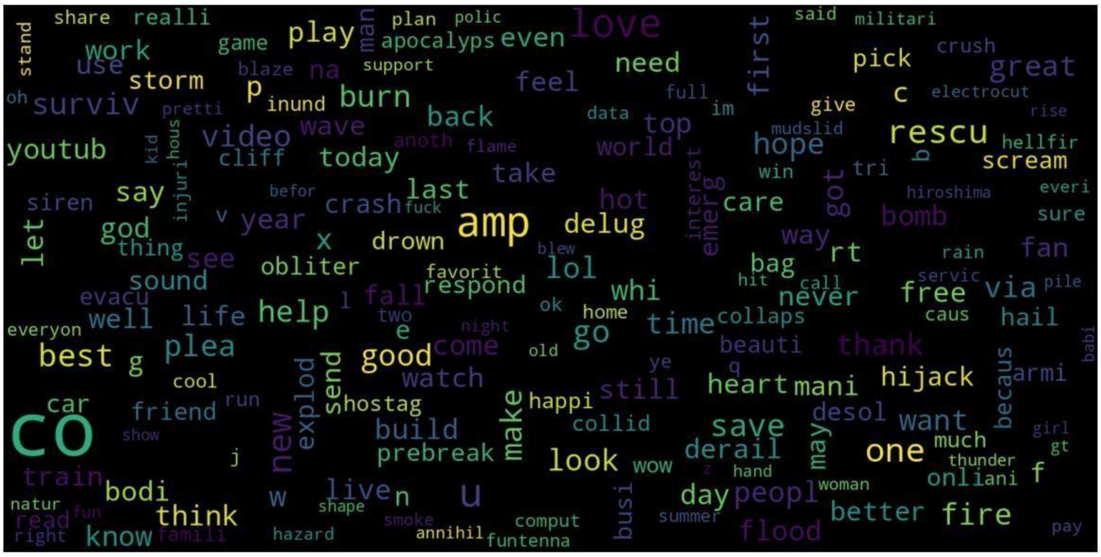
Plot used : barplot & Worldcloud

Result: The Above Bar plot represents the most frequent words in the reviews so that we can get a rough idea about the tweets. We can see that co is the most frequent word in tweets , most of the people types .com or c/o. Other frequent words that is used are like, http, fire. Variations starting with echo is more in use.

Most frequently occurring words in Disaster tweets

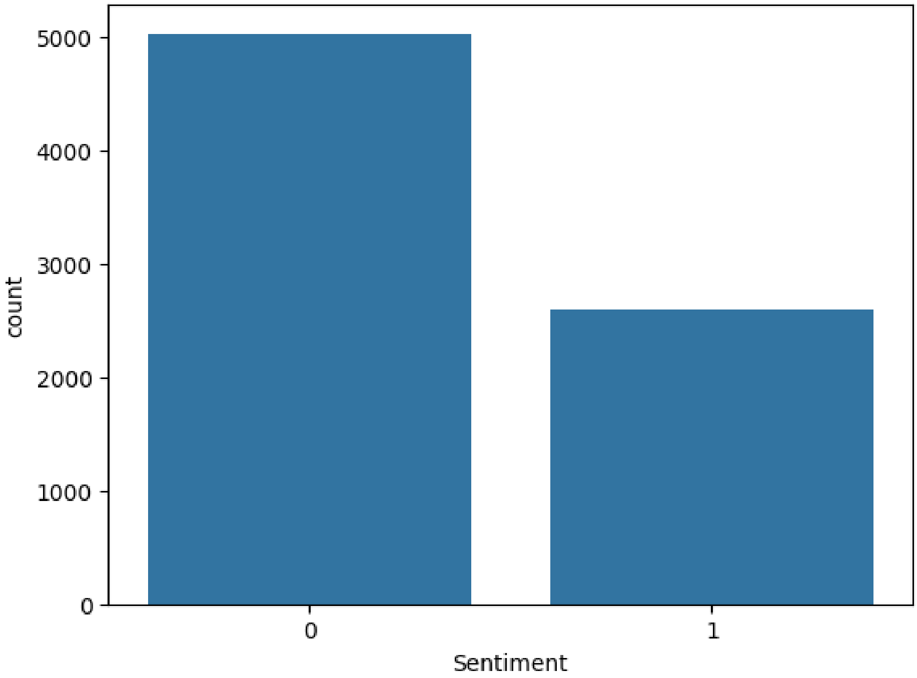


Most frequently occurring words in Not Disaster tweets



This word cloud tell us the most frequently occurring positive reviews. CO, fire, aattack is the word that appears most common in Disaster tweets. CO, amp, good is the word that appears commonly in not disaster tweets.

- **Visualizing target column–Sentiment**



Library : matplotlib & Seaborn

Plotused : Countplot

Result: The Above Count plot ,it can be observed that the total number of samples in Class 1 is around 2500 while in Class 0, it is about 5000. It means that positive tweets have around 2500 and negative tweets have around 5000

6. MODEL BUILDING

6. MODEL BUILDING

6.1. USE OF ALGORITHMS IN MODEL BUILDING

Complex models like ensembles and neural networks typically offer superior performance and accuracy. Despite their advantages, these models can be more challenging to interpret, making them less favorable. In the context of building classification models to predict social media disaster based on textual tweets, natural language processing (NLP) toolkits in Python prove indispensable. These toolkits enable data scientists to preprocess and analyze textdata effectively, extracting meaningful insights to inform decision-making processes. Thus, while more complex models may yield better performance, businesses must weigh the trade-offs between interpretabilityand accuracy when selecting the appropriate machine learning approach for their specific needs.

6.2. USE OF NLP IN MODEL BUILDING

To handle natural language, we use a tool called NLTK (Natural Language Toolkit) in Python.

Stop words

Natural Language Processing (NLP) in Python involves addressing various challenges, particularly in natural language understanding. Text data often includes common stop words like 'the', 'is', and 'are', which do not carry significant meaning and can be filtered out during processing. While there is no universal list of stop words in NLP research, the NLTK (Natural Language Toolkit) module provides a commonly used list. The initial step in NLP preprocessing involves removing stop words from the text to be analyzed. By eliminating stop words, NLP algorithms can better identify meaningful patterns and insights within the text data.

Corpus

A corpus refers to a comprehensive collection of written texts, encompassing either the complete works of a specific author or a body of writing centered around a particular subject. Within the Natural Language Toolkit (NLTK) package, various modules offer functions designed to read corpus files in diverse formats. These functions serve the purpose of extracting textual data from corpus files, facilitating analysis and processing. Notably, they can handle corpus files distributed within the NLTK corpus package, as well as those sourced from external corpora.

Snowball Stemmer

The Snowball Stemmer, also known as the Porter2 Stemmer, is a versatile tool capable of handling words from various languages, not limited to English. Renowned for its speed and efficiency, it excels in processing small segments of text with precision. Compared to the Porter Stemmer, it offers enhanced performance and accuracy, making it a preferred choice for stemming tasks.

6.3. DATAPROCESSING AND SPLITTING

Before splitting the dataset and constructing the model, the predictor variable (X) and the response variable (y) are defined. In this dataset, there is one column containing features and a total of 10,876 rows of data. The model is built for classification by utilizing X and y.

The dataset is divided into training and testing sets, with20%ofthe data allocated for testing the model and 80% for training the model. This split ensures that the model is trained ona substantial portion of the data to learn patterns and relationships effectively, while also allowing for an independent evaluation of its performance on unseen data.

Using the sklearn library, we easily split the dataset into two distinct sets: one containing the independent features (X) and the other containing the dependent variable (y). Subsequently, we further partition the X dataset into training (X_train) and testing (X_test) sets, while also splitting the y dataset into corresponding y_train and y_test sets. This straightforward process enables us to efficiently organize the data for training and evaluating machine learning models.

6.4. BUILDING MODELS

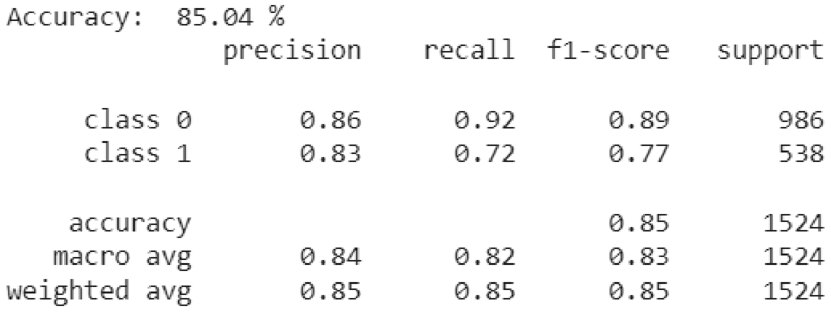
In this project, various algorithms from the Python ecosystem are used for data analysis, model creation, and prediction tasks. Some of the primary algorithms employed include:

- Logistic Regression
- Random Forest Classifier
- Decision Tree Classifier
- Support Vector Machines (SVM)
- Artificial Neural Networks (ANN)

To evaluate and select between models, we compare their performance using various metrics. Accuracy, calculates the ratio of correctly predicted observations to the total number of observations. Precision measures the ratio of correctly predicted positive observations to the total predicted positive observations, while recall quantifies the ratio of correctly predicted positive observations to all observations in the actual positive class.

Logistic Regression

Logistic regression, a classification algorithm, assigns observations to discrete classes, making it suitable for various classification tasks. Examples include distinguishing between email spam and legitimate emails. In this project, the logistic regression classification model is built using the logistic regression implementation from the linear_model module in the sklearn library. After training the model, key evaluation metrics such as the classification report and confusion matrix are generated to assess its performance and effectiveness in making accurate predictions.



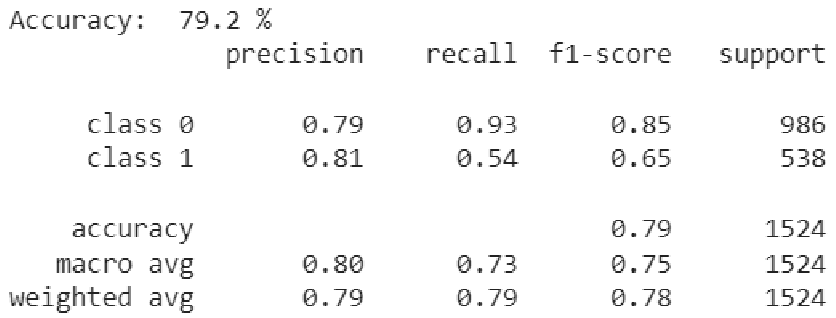
```
Accuracy:  85.04 %
              precision    recall  f1-score   support

     class 0       0.86      0.92      0.89       986
     class 1       0.83      0.72      0.77       538

    accuracy                           0.85      1524
   macro avg       0.84      0.82      0.83      1524
weighted avg       0.85      0.85      0.85      1524
```

Random Forest Classifier

Random forests, also known as random decision forests, are ensemble learning methods used for various tasks such as classification and regression. These methods operate by constructing numerous decision trees during training and then aggregating their outputs to make predictions. Specifically, the random forest algorithm outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. In this project, the classification model is built using the random forest classifier, which is loaded from the ensemble module in the sklearn library.



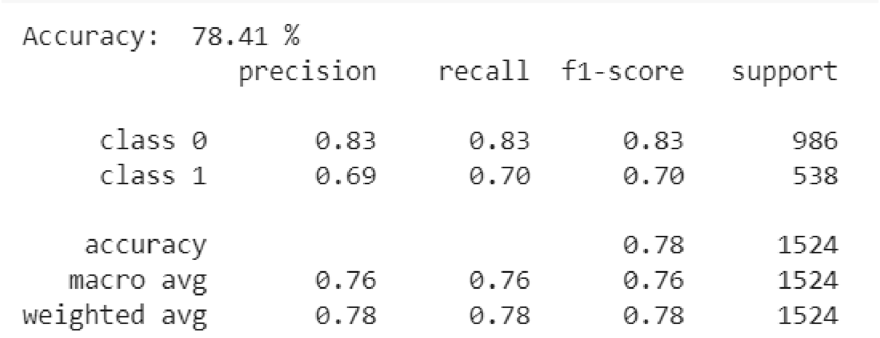
```
Accuracy:  79.2 %
              precision    recall  f1-score   support

     class 0       0.79      0.93      0.85       986
     class 1       0.81      0.54      0.65       538

    accuracy                           0.79      1524
   macro avg       0.80      0.73      0.75      1524
weighted avg       0.79      0.79      0.78      1524
```

Decision Tree Classifier

The Decision Tree Classifier is a widely used machine learning algorithm particularly effective for classification tasks like disease prediction. This algorithm constructs a decision tree by recursively splitting the dataset based on the most significant features until each leaf node represents a pure class or a small subset of the data. In the project, the classification model is built using the Decision Tree classifier, which is loaded from the tree module in the sklearn library. After training the model, the accuracy scores for both the training and test datasets are computed and provided below.



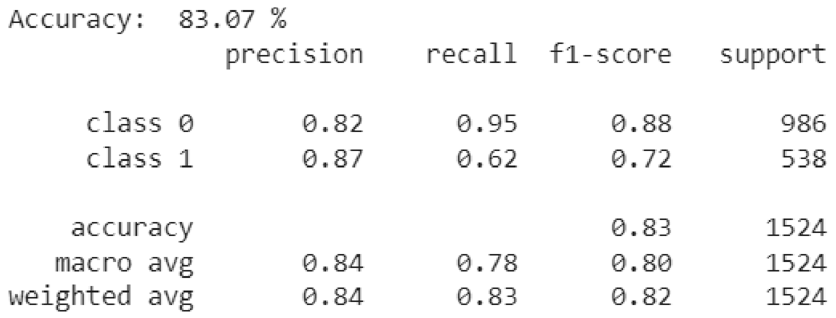
```
Accuracy:  78.41 %
              precision    recall  f1-score   support

     class 0       0.83      0.83      0.83       986
     class 1       0.69      0.70      0.70       538

    accuracy                           0.78      1524
   macro avg       0.76      0.76      0.76      1524
weighted avg       0.78      0.78      0.78      1524
```

Support Vector Machines (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm primarily utilized for classification tasks. It operates by representing each data point as a point in an n-dimensional space, where n represents the number of features. The algorithm then seeks to find a hyperplane that effectively separates the data points of different classes. This hyperplane acts as the decision boundary, enabling SVM to classify new data points based on their positions relative to this boundary.



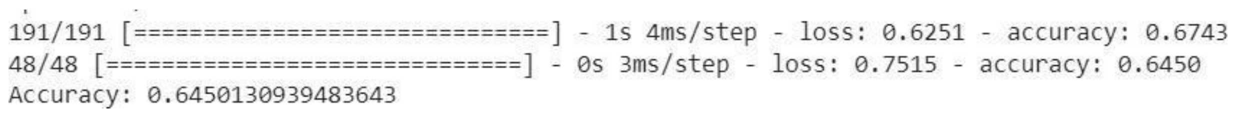
```
Accuracy:  83.07 %
              precision    recall  f1-score   support

     class 0       0.82      0.95      0.88       986
     class 1       0.87      0.62      0.72       538

    accuracy                           0.83      1524
   macro avg       0.84      0.78      0.80      1524
weighted avg       0.84      0.83      0.82      1524
```

Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) functions akin to the human brain, tackling challenging tasks such as pattern recognition and classification. Comprising interconnected layers of nodes known as neurons, each neuron receives, processes, and transmits information to subsequent layers. The connections between neurons possess weights that influence the network's output. Through training, the network adjusts these weights to improve its performance. ANNs excel at learning intricate patterns from data, making them well-suited for tasks like image recognition and natural language processing. However, training ANNs demands substantial computational resources and extensive datasets, and they can be more complex to comprehend compared to simpler models like decision trees.

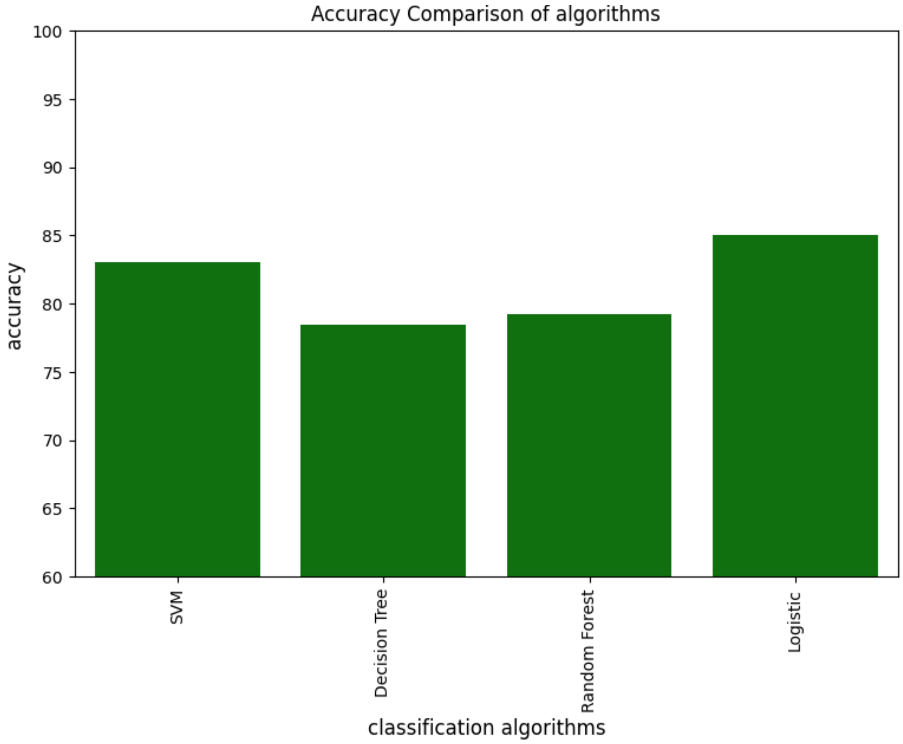


```
191/191 [==============================] - 1s 4ms/step - loss: 0.6251 - accuracy: 0.6743
48/48 [==============================] - 0s 3ms/step - loss: 0.7515 - accuracy: 0.6450
Accuracy: 0.6450130939483643
```

6.5. COMPARISION OF ALGORITHMS

From the above model building the accuracy score is compared as follows.

- Random Forest Classifier-79.2%
- Logistic regression-85.04%
- Support Vector Machine (SVM)-83.07%
- Decision Tree-78.41%
- Artificial Neural Networks (ANN)-64.5%



We chose to use an Logistic Regression with a high accuracy of 85.04% for building our model. This model provides better predictions. Now, we use the model to make predictions for new inputs.

7. FUTURE ENHANCEMENT

7. FUTURE ENHANCEMENT

In an era where social media platforms have become integral sources of information during disasters, enhancing the capabilities of analyzing and responding to social media disaster tweets is paramount. As such, future enhancements in this domain aim to use advanced technologies and methodologies to improve the detection, classification, and response to disaster-related tweets in real-time. These enhancements will not only enable more efficient monitoring and analysis but also facilitate timely and targeted interventions, ultimately contributing to more effective disaster management and response efforts.

- ❖ **Automated Detection and Classification:** Future enhancements will focus on developing advanced algorithms and machine learning models to automatically detect and classify disaster-related tweets in real-time. By leveraging cutting-edge natural language processing (NLP) techniques and deep learning architectures, such as transformer-based models, the system will be capable of accurately understanding tweet content and identifying relevant disaster events.
- ❖ **Mixed-media Analysis:** Enhancements will involve analysis techniques to analyze not only text but also images and videos shared on social media platforms during disasters. This approach will provide a more comprehensive understanding of unfolding events and enable responders to assess the situation more accurately.
- ❖ **User Verification and Credibility Assessment:** Implement techniques to authenticate and assess the trustworthiness of information shared on social media platforms during disasters. This involves deploying algorithms and strategies to verify the identities of users sharing information, evaluate the consistency and accuracy of their posts, and cross-reference data with reputable sources.

In conclusion, future enhancements in the field of social media disaster tweets aim to use advanced technologies and methodologies to enhance the detection, classification, and response to disaster-related events in real-time.

8. CONCLUSION

8. CONCLUSION

The Social Media Disaster Dataset, obtained from Data World, contains tweets concerning various disasters shared on social media platforms. In this project we tried to predict whether a tweet is an disaster tweet or not using the information from various social media platform. We tried difference models using Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Support Vector Machine

The aim of the project was to achieve a model with good accuracy and we got a model with an accuracyof85.04% , which is pretty good We got Logistic Regression as the best model with an approximate accuracy of 85.04%. Using outside data we evaluated model performance. The technology used is Python3 and editor used is Goggle Collab.

9. REFERENCES

9. REFERENCES

- https://data.world/crowdflower/disasters-on-social-media
- https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-stepbecd4d56c
- https://cognitiveclass.ai/badges/machine-learning-python/