DisasterTweetsEnsemble: Ensemble for Disaster Tweets Classification

MSc Research Project
DATA ANALYTICS (MSCDAD_JAN23C)

SHREEDHAR SHARMA
Student ID: 22142835

School of Computing
National College of Ireland

Supervisor: DR. ARGHIR NICOLAE MOLDOVAN

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	SHREEDHAR SHARMA		
Student ID:	22142835		
Programme:	MSc DATA ANALYTICS	**Year:**	2023-24
Module:	MSc RESEARCH PROJECT		
Supervisor:	DR. ARGHIR NICOLAE MOLDOVAN		
Submission Due Date:	31.01.2024		
Project Title:	DisasterTweetsEnsemble: Ensemble for Disaster Tweets Classification ……………………………………………………………………………………………………		
Word Count:	**6655**	**Page Count: 20**	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: SHREEDHAR SHARMA

Date: 30.01.2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

DisasterTweetsEnsemble: Ensemble for Disaster Tweets Classification

Shreedhar Sharma
x22142835

Abstract

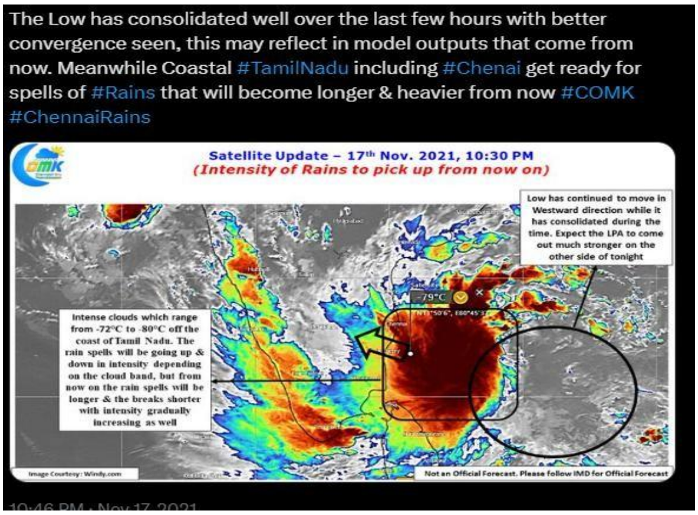
Analyzing social media data can significantly impact society, and one use case of social media is natural calamities. During natural calamities, one requires live feed from the disaster-affected area to provide aid, and social media is at the forefront of providing the data. The Research shows the use of social media Tweets during disasters to provide aid by filtering the data as relevant or non-relevant. At the time of the disaster, tweets or social media-generated data were high in velocity, and some random tweets that were not useful were also generated. Hence, the given Research tries to classify the tweets as informative and non-informative based on the labelled dataset. The Research trains various machine learning models on three datasets from CRISISNLP: Nepal, Queensland, and Crisis. The study uses the KDD methodology to work on the dataset, from cleaning the data with the help of natural language understanding to building different machine learning and deep learning algorithms. The study uses different vector techniques on the text data and builds models like Logistic Regression, Naïve Bayes, XGBoost and LSTM. The study shows that the count vectorizer performs well with different combinations of algorithms with a maximum accuracy of 95% on the Queensland dataset. Also, the study tries to build a state-of-the-art embedded model to go one step further in this direction and achieve an accuracy of 95.9 % on the Queensland dataset.

1 Introduction

1.1 Background

Social media is increasing quickly worldwide and becoming the most suitable communication method. The rise of social media apps where people can share thoughts and media growing, like Facebook and Twitter. According to the World Bank, 60 per cent [1]of the world uses social media, and there is no sign of a decline in the given trend. The given rise has several benefits as well as adverse effects. The following is used to capture sensitive information and public opinion worldwide. From modern wars like Russia and Ukraine to natural calamities, social media has proved its ability to provide accurate and up-to-date information. The information can be used to derive policies and send care and fast aid to these affected areas. The given Research studies the use of social media, especially Twitter data, in finding the relevant tweets regarding the disaster and filtering the irrelevant tweets based on the labelled dataset. The following will lead to a better arrangement of the resources and rescue aid by getting real-time information from many tweets. Our Research focused on building models on three different datasets to build general models that can process the data not based on disaster type. The novelty of the Research is to apply an embedded approach that compares its result with ensemble learning models.

.



The Low has consolidated well over the last few hours with better convergence seen, this may reflect in model outputs that come from now. Meanwhile Coastal #TamilNadu including #Chenai get ready for spells of #Rains that will become longer & heavier from now #COMK #ChennaiRains

Satellite Update – 17th Nov. 2021, 10:30 PM
(Intensity of Rains to pick up from now on)

Low has continued to move in Westward direction while it has consolidated during the time. Expect the LPA to come out much stronger on the other side of tonight

Intense clouds which range from -72°C to -80°C off the coast of Tamil Nadu. The rain spells will be going up & down in intensity depending on the cloud band, but from now on the rain spells will be longer & the breaks shorter with intensity gradually increasing as well

Image Courtesy: Windy.com

Not an Official Forecast. Please follow IMD for Official Forecast

10:46 PM · Nov 17, 2021

Figure 1 Disaster Tweet

1.2 Motivation

The Research focuses on building a classifier that can classify the relevant tweets so that one can first get valuable information during the disaster. The following study tries to pitch machine-learning algorithms to learn from the data and identify the tweets. In the Nepal disaster, most of the tweets were either in the local language or in Romanized words, and the following can create a problem in processing the tweets during the crisis. Hence, the study tries to build a model based on the domain knowledge and the usage of Natural language understanding (NLU) and NLP (Natural language processing). Applying various vector techniques and combinations of machine learning models to design a new algorithm that can embed various ML models inside it and provide maximum accuracy compared to singular machine learning and deep learning models is the main contribution in the given domain. The following will help provide help in real-time and get some awareness about the situation at the time of disaster. Also, the given data can be used to verify the information provided, which can help in the emergency response of the aids. The crisis team and the public both receive trustworthy updates.

1.3 Report Structure

The given Research is discussed in the report as follows:

a. **Introduction:** Chapter 1 introduces the problem of disaster tweet analysis, discusses the motivation behind the topic, and builds an embedded approach to solve the problem.
b. **Literature Review:** The next chapter shows the literature review on the given topic, focusing on finding the research gap and building a roadmap to solve the problem with novel results.
c. **Research Methodology:** Chapter three introduces the steps taken to solve the given problem based on the roadmap provided in Chapter Two and tries to implement the KDD methodology to classify the information as informative or not.
d. **Design Specification:** The chapter introduces the new approach to solving the given problem, stating the novel idea and proposed solution.
e. **Implementation:** The fifth chapter shows the methodology implementation on the text data and provides an in-depth analysis of the Research done.
f. **Evaluation:** The sixth chapter discusses the results obtained from the given Research and compares different models based on various evaluation metrics, discussing the final results.
g. **Discussion and Conclusion:** The last chapter provides the results obtained, discusses the results of the given problem, provides limitations and discusses all the plans

exploring the domains of machine learning and AI in the respective field and focusing more on end-to-end deployment of the model.

1.4 Research Questions

1. How well can the model behave in different types of disasters?
2. How can the model classify the relevant and non-relevant information based on the data?
3. How can the given analysis help in providing aid at the disaster location?
4. How accurately can these machine learning models predict the classes of text?
5. What is the impact of the given study on the organizations and policies?

1.5 Research Objectives

1. Collect the data and verify the quality of the data to build effective ensemble-based classification models.
2. Implement the natural language understanding and natural language processing to clean the tweets based on domain understanding.
3. Build various machine learning, deep learning and embedded models and try to increase their accuracy with different approaches.
4. Evaluate these models on various metrics like AUC-ROC and accuracy and find a way to explain the interpretability of the model.

2 Literature Survey

2.1 Related Work

Machine Learning-Based Approach

Social media tells the ground truth during a disaster and provides real-time information. Using these tweets, governments can take action and declare a disaster. (Lamsal & Kumar, 2020). The Research shows that working on a natural language processing-based task entirely differs from a data science project. The data collected from these tweets is analyzed in real-time, removing the noise from the data like URLs, numerics, digits, etc. The text is standardized, and all the stopwords are removed from the text, followed by the lemmatization. Once done, the features were extracted from the data via Unigrams, and to convert the data into a machine-understandable form, a Bag of Words and Word-count transformations were performed. The given model is then validated via K-fold stratified cross-validation, and the features are selected via the statistical approach of Chi-square. Once the data is ready, different classification models like Logistic Regression, Naïve Bayes, Random Forest, KNN, and ANN are built, comparing the machine learning algorithms with deep learning and are tested on various metrics. The logistic regression model outperformed all the other models in the given task with an accuracy of 73.82 per cent and an F-measure of 0.7615.

Social media data based on the source has a different significance. Twitter data that is shorthand and unstructured may have some bias in it. Hien et al., in their paper, have worked on a novel approach to identify disaster-related tweets by finding the common keywords, hashtags, etc. The Research uses data from various sources and Twitter, grouping them into a more extensive dataset based on each disaster type. The study utilized the geotagged tweets for the location access. For each region, the tweets are organized as relevant and irrelevant.

Once done, they are mapped as disaster or non-disastrous tweets, and the sentiment analysis is performed. To identify similar tweets, a machine learning approach is used in a refined manner. The second approach is based on learning where the data is tokenized, and with the help of a bag of words model, the data is classified as relevant and irrelevant via logistic regression. For the given study, the data is prepared by writers only, and it was concluded that the matching-based approach is high in quality compared to the machine-based techniques (To et al., 2017).

Medical resources are at their peak during a disaster, and different organizations need different types of information to provide the requirements. In their paper, (Madichetty & M, 2020) gave a new approach to [1]identifying the medical resources tweets using machine learning. Different Tweet IDs are collected via Twitter API, and the following tweets are tokenized and normalized. For the text cleaning part, the stop words and the numerical data from the tweets are removed using Regex. All the non-English comments are also removed from the data, and features are extracted based on the people's information, type of disaster and other related features. Once done, the data is used in different algorithms like support vector machine and bagging algorithms like Adaboost, Random Forest and Gradient boost. These models are compared with the state-of-the-art BoW model, and all the bagging classifiers and the vectorization techniques perform well.

Deep Learning-Based Approach

Rajesh Parsad et al., in their paper on the identification and classification of transportation disaster tweets, used BERT as the primary model. The Research focuses on reducing transportation-related accidents in Nigeria with the help of Twitter by identifying the location via Twitter API and the accident's severity. Previous Research on the given problem is based on the GPS data, but most people do not have the GPS enabled on their phones; hence, Named Entity Recognition is the only choice. The Research used real-time Twitter data from the Twitter API to classify whether the accident was disastrous. The given task is completed simultaneously with the help of the NER and Tweet via the location database. The shared tweets are cleaned and preprocessed via the natural language understanding, removing the non-English tweets, hyperlinks and other extra symbols via regular expressions. Also, to convert the data into an algorithm learning form, it is converted into word embeddings via Word2vec, a text metric. Different models like BERT, SVM, Random Forest, Decision Tree, and Xgboost are built once the data is ready. They are evaluated on the F1-score accuracy score, and the BERT model with AdamW beats all the other models. The proposed model increases the accuracy of the BERT via the AdamW optimizer. However, the following Research is only done for binary classification; hence, applying it to multi-class categories is a limitation (Prasad et al., 2023)

(Song & Huang, 2021) in their paper, have analyzed the real-time Twitter data for the prediction of real-time disasters. The given Research focused on collecting data from an Appen company with almost 10876 tweets, where 4962 are positive tweets and 6184 negative tweets. Hence, the data is balanced. The raw data from Twitter is cleaned by removing the stopwords, hashtags and emoticons and applying natural language understanding techniques. Once done, the given model is put up in a sequence of words and tokenized to generate the pipeline. The Research focuses on the learning pipeline of the SentiBERT that is used to transform words into tokens. The next model used in the study is BiLSTM to capture the order information and the sequence dependency. The CNN model is the last model used for the feature extraction in the given pipeline. The presented models are built via the combined binary cross entropy and focal loss. These models are then evaluated on precision, recall and F1 score to understand how good and evil the models predict the values. Given models are trained on different epochs and batch sizes where the max F1 score is 0.8956. The given model is good in finding the disaster tweets, but words having similar meanings, like apocalypse, are a drawback, and hence, the model can be fine-tuned by taking the same samples in both classes.

(Kabir & Madria, 2019) People use social media to communicate more frequently. Twitter allows users to share real-time information along with their location; hence, specific measures can be taken. The given research paper focuses on using the deep learning approach to classify tweets and build rescue scheduling. The Research collected data from two different hurricanes via Twitter Stream API. The given data is then preprocessed via various data cleaning techniques like discarding the non-English Tweets, removing duplicates, removing noise removal, and removing special characters and jargon. The given data is then fed into the deep neural network. The presented model consists of the input layer where the preprocessed data is provided, and then it goes to the embedding layer where pre-trained word vector models like GloVe are used. Next, the data will move to the LSTM layer to understand the long-term dependencies. The given data is then engineered to get the auxiliary features like the location and the other numeric values that can improve classification. Once these features are extracted, the model will feed into the convolutional layer, followed by the output layer. The location extraction is done via the metadata as the location data is unavailable. The priority scheduling system then uses the given data with multiprocessors to serve the people. The Research classifies the data into six categories and assists them with the help of a multi-task hybrid model. The given model is also compared with LR, SVM and CNN with the auxiliary features.

(Shekhar & Setty, 2015) The following paper discusses a natural disaster analysis interface that can find people's stress based on the tweets generated. The report researched the usage of many tweets available during the disaster. The data extracted from these tweets is via the different keywords and categorized into the following categories: Earthquakes, Floods, Forest Fires and Droughts. The dataset about each class is created via the K-nearest neighbour. Tweets generally used on Twitter are of shorter length and contain specific acronyms. Hence, the given system uses certain techniques to find the redundant tweets and preprocess it effectively. The shared tweets are categorized into different categories based on the keywords available in the dataset. Next, the analysis performed on the given dataset is based on Geo-Location tagging, and the location of a particular tweet is mapped via the Google API. Based on the site, they are clustered in a specific area and analyzed over time. Once the final tweets are available, they are fed into the sentiment analysis program, and based on the level of negativity, the tweet is categorized into different categories.

Akash Kumar et al., in their paper, have analyzed the CrisisMMD- a novel dataset for developing crisis response mechanisms. The report utilizes the given dataset on different calamities. The delivered dataset contains text and images, and ImageNet is a prebuilt architecture. The clean dataset built over the given dataset is fed into the transfer learning models with the softmax function to get the final output. The text data is provided in the text-only models. First, the data is cleaned per the natural language understanding, removing different stopwords and tweets with lengths of less than three words. The given model is then fed into the N-gram model, and other ML models are built on top of that. Following deep learning models are used for ternary classification. LSTM, Bi-LSTM and CNN with Glove are used with pre-trained word embeddings. For the image dataset, the best accuracy is achieved by Inception and for text, the Random Forest model outperforms all other models. Combining both the image and text-based models, the logistic regression performed well. The given model can be used to find the information and provide emergency aid in these disasters (Gautam et al., 2019).

(Madichetty & Sridevi, 2021) In their new paper, the same authors used a neural network-based approach to detect the information regarding a disaster using the Twitter dataset. The following situational awareness data can help authorities understand the data and the situation in a better way. The Research provides a feature-based approach to detect tweets in both Hindi and English. The data here can be from situational or non-situational tweets and hence needs to be understood. The Research focuses on implementing the RoBERTa on the feature-based approach with the fusion of multiplicative vectors to get the final results. The data is case-folded and cleaned via POS tagging and lemmatization, and all the unnecessary symbols are removed from the dataset. The given data is then extracted with the help of personal pronouns and other features to classify the tweets as situational or non-situational using SVM. The given data is then applied with the help of the RoBERTa model, as the following model has an advantage over BERT. These models are then compared with the baseline models, and the next model received a precision score of 95-98 per cent based on different settings.

Descriptive Analysis

Social media in recent years has shown so much growth. Nowadays, people express their views over social media through different platforms. The authorities can use this information as an opportunity to extract useful information. However, disaster-related analytics methods are not versatile enough to get the ground report. Hence, Nayomi et al., in their paper, used a well-established data analytics framework to work on real-time Twitter data. The given study is done on the Southeast Queensland (SEQ). The presented Research uses the Capture-Understand-Present (CUP) framework, a three-stage process, and a data analysis process to find the severity level of the disaster and the spatial dimension sentiment analysis to find the high-impact areas. The data is captured via Twitter API, and the data is processed with various filters. The data is removed from multiple URLs, then the content analysis is performed to get the meaningful content out, and different machine learning algorithms are applied. Secondly, text clustering is performed to get the clusters out and find the basic needs. The data here is geotagged; hence, the study uses inverse distance weightage (IDW) to interpolate the disaster impact on various areas (T., Goonetilleke, A., & Kamruzzaman,2019)

Year	*Approach Used*	*Authors*	*DataSet*
2020	Logisitic Regression,Naïve Bayes,Random Forest, KNN and ANN	R. Lamsal and T. V. V. Kumar	Data by Figure Eight Inc.
2017	Matching Based Learning/Learning-Based Learning	H. To, S. Agrawal, S. H. Kim, and C. Shahabi	CrisisLexT6 - Wildfire
2020	Adaboost, Random Forest and Gradient boost	S. Madichetty and S. M,	Fire 2016, SMERP2017
2023	BERT, SVM, Random Forest, Decision tree, and Xgboost	R. Prasad, A. U. Udeme, S. Misra, and H. Bisallah	Transportation Dataset via Twitter API
2021	CNN, BiLSTM, SentiBERT	G. Song and D. Huang,	Disaster data (Nepal and India) via Twitter
2019	Deep Learning model with seven components mainly: BLSTM, CNN, etc.	M. Y. Kabir and S. Madria,	Hurricane Harvey, Hurricane Irma, CrisisNLp, CrisisLex
2015	Integrated NLP with KNN	H. Shekhar and S. SettyDisaster	Earthquake Dataset – Via Twitter
2019	VGG-19, Decision Tree	A. K. Gautam, L.	Crisis MMD

		Misra, A. Kumar, K. Misra, S. Aggarwal, and R. R. Shah,	
2020	CNN, LSTM, BiLSTM	S. Madichetty and M. Sridevi	Nepal Earthquake, SanHshoot
2019	Content Analysis, Descriptive Analysis	Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, M.	Nepal Earthquake and Others

Table 1 Literature Comparison

2.2 Summary and Discussion

Out of all the papers discussed, the important thing is to find the Geo-location of the tweets as provided by the Twitter metadata or by tagging the location. The following is the main thing while analyzing the severity of the disaster. The next thing is that with the help of descriptive analytics, it is possible to find the number of tweets in a region. Then we can extract the critical hashtags or topics to find the essential hashtags that can be used to filter the data. While working with the tweets one should follow the natural language understanding process. Also, people are using content analysis, machine learning and deep learning to classify the tweets.

Now, the main roadmap that is coming out of the given Research is that none of them have implemented any data mining technique. Hence, CRISP-DM, SEMMA, or KDD can be used to make the whole implementation much more accessible and also, with the help of natural language understanding and processing, a comparative study between descriptive analytics, machine learning and deep learning can be discussed.

3 Research Methodology

3.1 KDD:

KDD is a standard data mining process which is used in data mining. However, the following can also be used in text mining, known as Knowledge Discovery from Text (KDT).[12][13]The working of the given methodology on the text data differs from KDD as here, along with analytical functions, functions from the Natural Language are also applied. The following methodology is a seven-step process used along with text mining in our project.

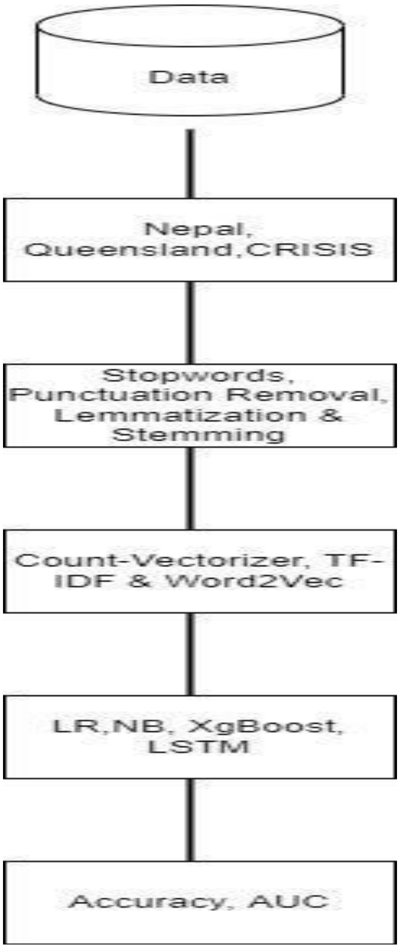

Data
Nepal, Queensland,CRISIS
Stopwords, Punctuation Removal, Lemmatization & Stemming
Count-Vectorizer, TF-IDF & Word2Vec
LR,NB, XgBoost, LSTM
Accuracy, AUC

Figure 2 KDD Methodology

3.1.1 Data Selection:

The first step in any data-based project is to get the data. The data used in our Research is based on the disaster tweets dataset and is collected from CRISISNLP. Data selection is a process where the data quality is discussed, and the credibility of the data sources is discussed. The data used in our Research is from a verified source; all the citations are available in the link. The following dataset used in our Research is from different parts of the world. The first dataset is the Nepal Earthquake dataset, Queensland Floods and the CRISIS dataset, which combines multiple disasters worldwide. This dataset contains the text data, other attributes, and target column, whether the given tweet is informative. The main reason for choosing these datasets is to understand the machine learning models from a broader perspective and check the performance and best practices to be applied in disaster datasets. The dataset covers different types of disasters, and the study will show an overall analysis.

3.2 Data Preprocessing:

Once the data is available, it needs to be cleaned to raise the data quality for modelling. All the datasets have information about the Tweet-Id and other information that is not needed. The Research is concerned with the Tweet text and the class label. Hence, first, the data is subsetted as per that. Upon verifying the dataset, the class label is balanced and does not require further clarity. The primary thing is to clean the dataset, and for that, Natural Language Understanding(NLU) is taken in. The following helps identify the sentence's meaning, clean it semantically, and build the relevant text data per the business problem. The NLU applied on the dataset is as follows:

- **Case Standardization:** The data available in the Research is Twitter data containing multiple tweets comprising hashtags and other punctuations. First, the whole data is converted into lowercase, and the following will help in better data parsing.

- **Remove Punctuations:** Once the data is in lowercase, punctuations are removed from the dataset, and the most common way used for that is Regular expressions. These unique expressions will help identify and substitute the text strings as per command. The Research removed all the punctuation and special characters with the help of regular expressions.

- **Stop word Removal:** A corpus comprises characters, words and paragraphs, and according to Zipf's law, the frequency of a word distribution and the rank of the word in a large text document is inversely proportional.

$$\text{word frequency} \propto \frac{1}{\text{word rank}}.$$

Figure 3 Zipf's Law

Hence, one can conclude from the above law that the corpus comprises stop, significant, and everyday words. Hence, the main thing needed for analysis is regular and significant words. Rest all the words that need to be removed. The Research removed stop words based on the English language from the whole text and focused on the essential words.

- **Lemmatization and Stemming:** After removing the stop words, the data consists of all the everyday words, but there are chances that the people have used different forms of the words, which need to be treated. The next step is to use Stemming and lemmatization, where the words are returned to the base word. This will help get a better word meaning, which will help in better results after parsing.

3.3 Data Transformations:

After the data preprocessing, the data is ingested into machine learning models. However, these models cannot understand the data in raw text form and need to be converted into numbers. The concept used here is word embeddings, where the data is converted into numbers. All the different types of word embeddings are discussed here.

- **Count-Vectorizer:** The Count-vectorizer model converts the text data into numbers by counting them and putting them in vectorized form. The given approach has shown evidence of state-of-the-art results in analyzing disaster data, mainly social media data.[14],[15]

- **TF-IDF:** The TF-IDF vectorizer counts the number of words and the number of times that word is contained in a document in the corpus, providing the importance of words. The given metric is much better than the Count vectorizer as it puts more force on the words in the sentence, making the matrix better [16]. The following is used in studies where Twitter data is used for disaster analysis.

- **Word2Vec:** The given model is also a vectorized model where the model learns about the text and provides the text based on the learning by putting the data in higher dimensional space. The reason for using the given matrix is that in micro-blogging sites, there is much noise in the data, and the words are distributed in the vector space.

3.4 Data Mining:

Data mining is a technique where functional patterns are found out of data. However, the term is text mining, where the patterns are drawn from the text data. Different text mining techniques are used to collect patterns from the data. The given methods provide the best results and are finalized after studying the literature on the given problem. These methods have shown state-of-the-art results in their respective domain of NLP. The three machine learning algorithms Logistic Regression, Naïve Bayes and Xgbbost are used because of their better performance. Logistic regression is the simplest model; Naïve Bayes is based on probability, word distribution, and Xgboost because of high computational power. The last model used in the Research is LSTM, which is primarily used to process text data.

- **Logistic Regression:** Logistic regression is based on the logit statistical model. The one used in our analysis is the machine learning-based model, which gives the probability of an event or not based on the given dataset of independent variables. The given algorithm provides the output as 0 and 1. The main reason to use the given algorithm is its simplicity and the results. The following provides better results in a sparse matrix as it can easily divide the data into classes[17].

- **Naïve Bayes:** The next model used in the Research is the Bayes model, which is based on conditional probability. The following model depends on Baye's theorem, where every feature is independent. As the features based on which the classification happens are independent, it is called Naïve and Bayes based on Baye's Theorem [18]. The following algorithm can quickly capture the word patterns in the spars matrix and is used in the Research.

- **XgBoost:** The third algorithm used in the Research is XgBoost. Famous for its high performance. The algorithm pushes the limits of the computer to an extent that can be used for calculation in the boosted trees as the data will be in high-dimension space. That is the reason it is used here [19].

- **LSTM:** The last model used in the Research is LSTM, also known as Long Short-Term Memory. The given model is a recurrent neural network with a memory unit. The following model can behave better than most of the neural networks on the text data. as the given model is designed for the sequence data. The text data is always in sequence; hence, it can remember the last word and do the classification better [20],[21].

3.5 Evaluation:

Once the model is ready, we have to evaluate it based on the different metrics to justify the working of the model as well as understand how better it fits the objective. The Research uses different metrics, as one cannot rely on the accuracy of the models in general.

- **Confusion-Matrix**: The first matrix used is a confusion matrix of the number of correct classifications and the incorrect classifications given by True Positive, True Negative, False Positive and False Negative. The given matrix shows the number of misclassifications and accurate predictions by model in numbers, and one can easily visually see what happened.

- **ROC-AUC:** The second matrix used in the model is the AUC_ROC score. This matrix gives an accurate picture of the model and tells how good the model is in identifying the positive and negative classes. The given matrix is a combination of sensitivity and 1-specificity, which means that a true positive is identified as a true positive and how well the misclassifications are happening. The given matrix with a higher score tells that the model is better in classification.

- **Classification Report:** To sum up these two matrices, the other matrix used in the Research is the Classification report that tells the precision of the model, recall, support and F1-score. The given report can easily distinguish between the models when multiple models are used to show the comparisons and find the best model.

4 Design Specification

4.1 Proposed Solution:

The Research mainly focuses on building an embedded model that can classify the disastrous tweets as informative or not out of all the tweets on various datasets. Hence, the given solution provides the details of our solution. The data selected in the Research is from different datasets to test the models and explain the models' generalization.

1. The Research deals with highly unstructured text data, so the text's cleaning process differs from the standard data science project. Using the Natural Language Understanding (NLU), the data is cleaned per the above-defined transformations. Some significant steps are removing the stop words, converting the text into a lower format, and removing all the unnecessary punctuation. Lemmatization and stemming are used to make the data syntactically correct. The output obtained after applying all these steps is given as follows.

Dataset	Data Before Cleaning	Data After Cleaning
CRISIS	Organization that are working in Haiti, I do not have any ways of taking care of my family here	organization working haiti ways taking care famili'
	Maria now a hurricane again!! Strong storm surge is	maria hurricane strong storm surge issue

	an issue in #kitryhawk https://t.co/DSE2C5faKp https://t.co/vR63N0YZRY	kitryhawk httpstcodsecfakp httpstcovrnyzri
Nepal Floods	RT @AnupKaphle: #Nepal's prime minister addressed the country for 1st time since earthquake on Saturday. No concrete plans, lots of referenâ\x80°Ã\x9b_	rt anupkaphle nepals prime minister addressed country st time since earthquake saturday concrete plans lots referen
	@jonsnowC4 So have we; read our friends blog from Lamjung where they are working http://t.co/jGpSacUQpe	jonsnowc read friends blog lamjung working httptcojgpsacuqp
Queensland Floods	I just though about the night I went clubbing with @mikal1988 and I cried	though night went clubbing mikal cri
	Looks like its going to be another long night courtesy #djoko and #murray #AustralianOpen	looks like going another long night courtesy djoko murray australianopen

Table 2 Before and After comparison of data

2. The given table shows the cleaned and uncleaned dataset and one can observe that the Queensland data is cleaned very well compared to the other two samples when taken randomly. Also, we can see irrelevant tweets related to the Australia Open in Queensland dataset.

3. Once the dataset is ready, it should feed into the machine learning models; for that, all the different text metrics are used to convert the text data into numbers. Here, various combinations are applied. The data is converted into the Count-Vectorizer, TF-IDF and Word2Vec matrix, which are given as input to the machine learning models.

4. The data is then provided to various machine learning and deep learning models, and all the resultant combinations are discussed.

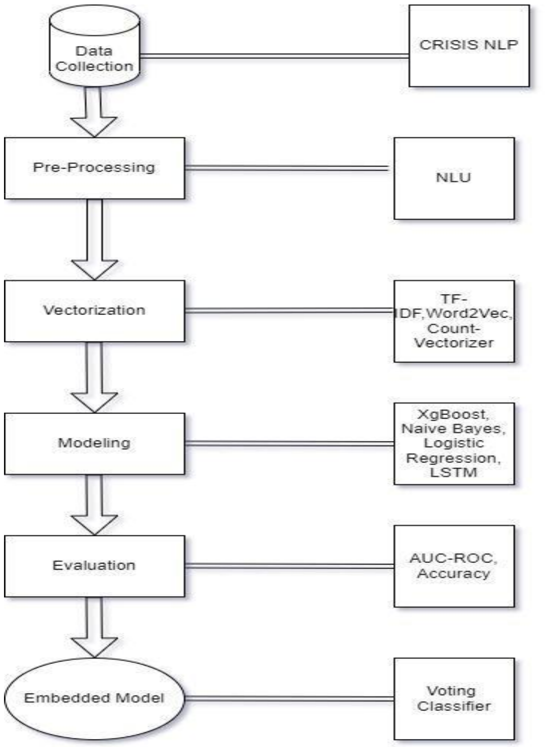


Figure 4 Flow Chart of research flow

4.2 Embedded Model:

The study has shown a novel approach to building an embedded model of top performers. The Research has shown the use of various machine learning models to build a complex voting classifier that can filter the tweets based on the data. The study is addressed after building various machine learning and deep learning algorithms. The approach showcases that machine learning algorithms perform well on text data when data is cleaned and vectorized best.

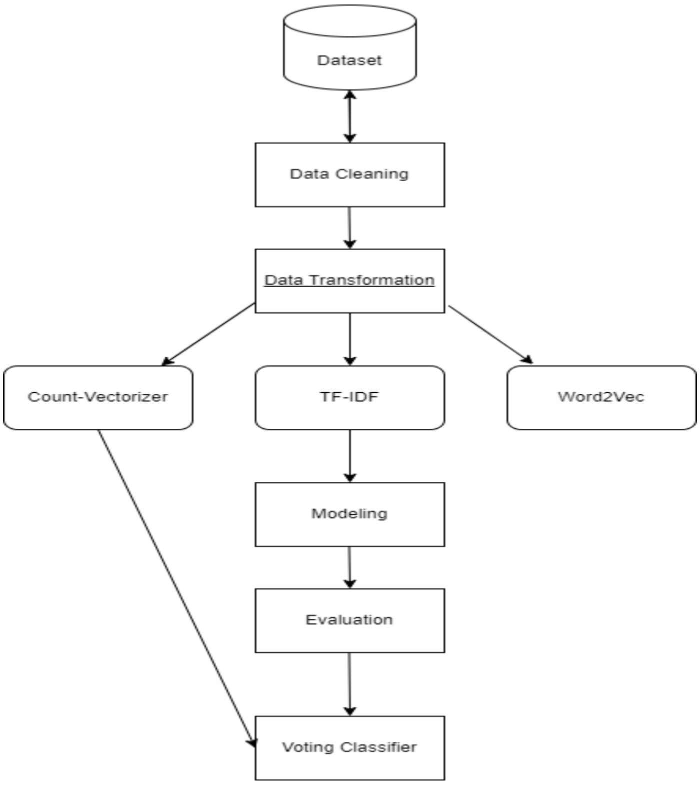


Figure 5 Embedded Model

5 Implementation

5.1 Tools and Technologies:

The whole study is implemented on Python Programming, and the tool used here is Jupyter Notebook. Python programming is used instead of R or MATLAB because Python (3.8) offers a wide range of libraries and open-source communities that can be used to solve various data science problems. Also, Python supports text analytics with libraries like Natural Language Toolkit (NLTK) and Spacy. Various other natural language data handling libraries are used in the Research to deal with the text data. Also, the given library provides a better visualization library that can capture every pattern in the data. Most of the code done on the Research is done on the Jupyter Notebook IDE. That supports the coding and the embedding of various pictures in the notebook. Also, the following has a cloud-based alternative, Google Colab, which is used in building the deep learning models in the Research. The versions of major libraries are shown below:

Libraries	**Version**
Pandas	1.4.4
Numpy	1.23.5
NLTK	3.7
Sklearn	1.0.2
Keras	2.12.0

Table 3 Libraries Version

6 Evaluation

The Research's main objective is to build a classifier that can classify the tweets as Relevant and Non-relevant on the text data, identifying the behaviour on different datasets. It also identifies the accuracy of the different models in making the prediction. The described models are performed on the different datasets along with the combination of various metrics, and the results are discussed as follows.

Nepal Dataset

The first dataset used in the Research is the Nepal dataset, and the results obtained from the Nepal dataset based on the accuracy and AUC_ROC score are as follows. The results show that the maximum accuracy obtained in the Nepal dataset is with a count-vectorizer matrix with various models except LSTM. As the data is highly dimensional, LSTM cannot perform better there. However, it showed better accuracy with the TF-IDF model. However, on verifying the results from the AUC score, the TF-IDF and count vectorizer show comparable results on these two matrices.

Model	Vector Matrix	Accuracy
Logistic Regression	Count-Vectorizer	0.772
	TF-IDF	0.773
	Word2Vec	0.526
Naïve Bayes	Count-Vectorizer	0.770
	TF-IDF	0.761
	Word2Vec	0.495
XgBoost	Count-Vectorizer	0.760
	TF-IDF	0.745
	Word2Vec	0.525
LSTM	Count-Vectorizer	0.497
	TF-IDF	0.773
	Word2Vec	0.524

Table 4 Accuracy of models on the Nepal Dataset

Model	Vector Matrix	AUC
Logistic Regression	Count-Vectorizer	0.769
	TF-IDF	0.772
	Word2Vec	0.503
Naïve Bayes	Count-Vectorizer	0.770
	TF-IDF	0.760
	Word2Vec	0.516
XgBoost	Count-Vectorizer	0.757
	TF-IDF	0.743
	Word2Vec	0.502
LSTM	Count-Vectorizer	0.490
	TF-IDF	0.846
	Word2Vec	0.519

Table 5 AUC score of models on the Nepal Dataset

Queensland Dataset

The next test case used in the Research is the Queensland dataset, where the same models have shown state-of-the-art results. In the given test case also the models have shown the best accuracy with the Count-vectorizer. With a maximum accuracy of 95.8 per cent with XGBoost, the same pattern is also observed in the AUC_ROC score.

Model	Vector Matrix	Accuracy
Logistic Regression	Count-Vectorizer	0.957
	TF-IDF	0.953
	Word2Vec	0.508
Naïve Bayes	Count-Vectorizer	0.941
	TF-IDF	0.934
	Word2Vec	0.510
XgBoost	Count-Vectorizer	0.958
	TF-IDF	0.955
	Word2Vec	0.509
LSTM	Count-Vectorizer	0.497
	TF-IDF	0.503
	Word2Vec	0.516

Table 6 Accuracy of models on the Queensland Dataset

Model	Vector Matrix	AUC
Logistic Regression	Count-Vectorizer	0.957
	TF-IDF	0.952
	Word2Vec	0.521
Naïve Bayes	Count-Vectorizer	0.942
	TF-IDF	0.934
	Word2Vec	0.523
XgBoost	Count-Vectorizer	0.958
	TF-IDF	0.955
	Word2Vec	0.523
LSTM	Count-Vectorizer	0.494
	TF-IDF	0.493
	Word2Vec	0.525

Table 7 AUC score of models on the Queensland Dataset

Crisis Dataset

The last dataset under consideration is the Crisis dataset. Here, also the count vectorizer matrix model has performed well compared to TF-IDF and Word2Vec. The TF-IDF, however, has shown better accuracy with machine learning and deep learning models. Because of the larger data in the Crisis dataset, the model may have learned the pattern better.

Model	Vector Matrix	Accuracy
Logistic Regression	Count-Vectorizer	0.861
	TF-IDF	0.852
	Word2Vec	0.605
Naïve Bayes	Count-Vectorizer	0.828
	TF-IDF	0.774
	Word2Vec	0.600
XgBoost	Count-Vectorizer	0.846
	TF-IDF	0.839
	Word2Vec	0.605
LSTM	Count-Vectorizer	0.815
	TF-IDF	0.598
	Word2Vec	0.604

Table 8 Accuracy of models on the Crisis Dataset

Model	Vector Matrix	AUC
Logistic Regression	Count-Vectorizer	0.857
	TF-IDF	0.837
	Word2Vec	0.510
Naïve Bayes	Count-Vectorizer	0.802
	TF-IDF	0.725
	Word2Vec	0.511
XgBoost	Count-Vectorizer	0.845
	TF-IDF	0.840
	Word2Vec	0.509
LSTM	Count-Vectorizer	0.513
	TF-IDF	0.493
	Word2Vec	0.889

Table 9 AUC score of models on the Crisis Dataset

Once these models are evaluated on the given metrics, the last model, the embedded model, is built out of the analysis of these models. The following model combines all the weak learners in terms of Ensemble learning. However, these models combine all the high-performing models achieved on the Count-Vectorizer. The last model is a hard voting classifier with the combination of Logistic Regression, Naïve Bayes and XgBoost, and the maximum accuracy achieved on the Queensland dataset.

Hence, the study concludes that the count vectorizer performed well on the disaster dataset, and the best accuracy obtained is on the Queensland dataset. Maybe the data is cleaned better than the other datasets, or the dataset size is small compared to the different algorithms. However, the Research done in the given field is also based on the domain and the type of text being used. However, the deep learning models have not performed well on the given dataset, but there is evidence of better performance. The Research tries to explore the given idea in the future by examining different datasets and data cleaning techniques and, again, trying to build a different combination that can be deployed in the real world for better analysis and to provide relief at the time of disaster. Also as compared to the previous research[22], the results obtained in the previous research have used the CNN model on the text data and received an accuracy of 0.80 in unimodal architecture.

Dataset	Accuracy
Nepal	0.779
Queensland	0.959
Crisis	0.859

Table 10 Accuracy of Embedded model on different datasets

7 Conclusion and Future Work

The Research has presented an embedded technique for analyzing Twitter data based on the disasters. The study showcased the application of natural language processing in identifying whether text tweets are informative. The study focuses mainly on CRISIS, the Nepal flood and Queensland datasets. The study has shown the use of KDD as a methodology for extracting knowledge from the text and has demonstrated the implementation of Natural Language Understanding on the three datasets. From removing stopwords to data cleaning, all the steps are the same. However, the results obtained differ based on the data type available.

The central theme is implementing the steps on the dataset and trying out the different vector methods. The study has shown various permutations and combinations of algorithms and models, and it is concluded that the Count vectorizer performs well on the test dataset, with maximum accuracy obtained at 95.8 per cent on the Queensland dataset with XgBoost. The given model has shown its simplicity in the binary classification. Once the analysis is done, an embedded model with a top classifier is built to increase the accuracy and answer the research approach. However, the study has worked on only the label dataset. In the future, we will try to use the live feed from social media using the APIs, and the image and text data should be used to get real-time analysis. One can identify fake tweets using various deep learning and machine learning models, and the following calculation can be used to get better results based on the location. Also, for future work, the study will try to deploy the model over the cloud, showing a proper state-of-the-art model in the implementation.

8 References

[1] https://data.worldbank.org/indicator/IT.NET.USER.ZS - Accessed on 12 Nov 2023

[2] R. Lamsal and T. V. V. Kumar, “Classifying Emergency Tweets for Disaster Response,” *International Journal of Disaster Response and Emergency Management*, vol. 3, no. 1, pp. 14–29, Jun. 2020, doi: 10.4018/ijdrem.2020010102.

[3] H. To, S. Agrawal, S. H. Kim, and C. Shahabi, "On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based," in *Proceedings - 2017 IEEE 3rd International Conference on Multimedia Big Data, BigMM 2017*, Institute of Electrical and Electronics Engineers Inc., Jun. 2017, pp. 330–337. doi: 10.1109/BigMM.2017.82.

[4] S. Madichetty and S. M, "Identification of medical resource tweets using Majority Voting-based Ensemble during the disaster," *Soc Netw Anal Min*, vol. 10, no. 1, Dec. 2020, doi: 10.1007/s13278-020-00679-y.

[5] R. Prasad, A. U. Udeme, S. Misra, and H. Bisallah, "Identification and classification of transportation disaster tweets using improved bidirectional encoder representations

from transformers," *International Journal of Information Management Data Insights*, vol. 3, no. 1, Apr. 2023, doi: 10.1016/j.jjimei.2023.100154.

[6] G. Song and D. Huang, "A sentiment-aware contextual model for realtime disaster prediction using twitter data," *Future Internet*, vol. 13, no. 7, Jul. 2021, doi: 10.3390/fi13070163.

[7] M. Y. Kabir and S. Madria, "A deep learning approach for tweet classification and rescue scheduling for effective disaster management," in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, Association for Computing Machinery, Nov. 2019, pp. 269–278. doi: 10.1145/3347146.3359097.

[8] H. Shekhar and S. Setty, "Disaster analysis through tweets," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI),* 2015, pp. 1719–1723. doi: 10.1109/ICACCI.2015.7275861.

[9] A. K. Gautam, L. Misra, A. Kumar, K. Misra, S. Aggarwal, and R. R. Shah, "Multimodal Analysis of Disaster Tweets," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 94–103. doi 10.1109/BigMM.2019.00-38.

[10] H. To, S. Agrawal, S. H. Kim, and C. Shahabi, "On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based," in *Proceedings - 2017 IEEE 3rd International Conference on Multimedia Big Data, BigMM 2017*, Institute of Electrical and Electronics Engineers Inc., Jun. 2017, pp. 330–337. doi: 10.1109/BigMM.2017.82.

[11] M. Yehia, L. Fattouh, and M. Abulkhair, "Text Mining and Knowledge Discovery from Big Data: Challenges and Promise," Dec. 2019, doi: 10.20943/01201603.5461.

[12] A. Wilcox and G. Hripcsak, "Knowledge Discovery and Data Mining to Assist Natural Language Understanding."

[13] S. Soderland, "Learning to Extract Text-based Information from the World Wide Web Information Extraction from the Web," 1997. [Online]. Available: www.aaai.org

[14] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, "Using Social Media to Enhance Emergency Situation Awareness," *IEEE Intell Syst*, vol. 27, no. 6, pp. 52–59, 2012, doi: 10.1109/MIS.2012.6.

[15] R. Goyal, "Evaluation of rule-based, CountVectorizer, and Word2Vec machine learning models for tweet analysis to improve disaster relief," in *2021 IEEE Global Humanitarian Technology Conference (GHTC)*, 2021, pp. 16–19. doi: 10.1109/GHTC53159.2021.9612486.

[16] G. A. Dalaorao, "Applying Modified TF-IDF with Collocation in Classifying Disaster-Related Tweets," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 1.1 S I, pp. 28–33, Feb. 2020, doi: 10.30534/ijatcse/2020/0691.12020.

[17] B. Vimal, "Application of Logistic Regression in Natural Language Processing," International Journal of Engineering Research and, vol. V9, Dec. 2020, doi: 10.17577/IJERTV9IS060095.

[18] D. Phuc and N. T. K. Phung, “Using Naïve Bayes Model and Natural Language Processing for Classifying Messages on Online Forum,” in 2007 IEEE International Conference on Research, Innovation and Vision for the Future, 2007, pp. 247–252. doi: 10.1109/RIVF.2007.369164.

[19] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Dec. 2016.

[20] L. Yao and Y. Guan, “An Improved LSTM Structure for Natural Language Processing,” in 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), 2018, pp. 565–569. doi: 10.1109/IICSPI.2018.8690387.

[21] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, “Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism,” Applied Sciences, vol. 10, no. 17, 2020, doi: 10.3390/app10175841.

[22] G. A. Dalaorao, “Applying Modified TF-IDF with Collocation in Classifying Disaster-Related Tweets,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.1 S I, pp. 28–33, Feb. 2020, doi: 10.30534/ijatcse/2020/0691.12020.