
Texas A&M University - Commerce

Department of Computer Science

# Machine Learning for Disaster Detection through Twitter Analysis

Sneha Perithambi

*Supervisor:* Derek Harter, Ph.D.

A report submitted in partial fulfilment of the requirements of
Texas A&M University - Commerce for the degree of
Master of Science in *Computer Science*

April 17, 2024

## Declaration

I, Firstname(s) Lastname, of the Department of Computer Science, Texas A&M University - Commerce, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of TAMUC and public with interest in teaching, learning and research.

Sneha Perithambi
April 17, 2024

# Abstract

Twitter has become a crucial medium for people to use their smartphones to provide real-time views of events in the context of modern disaster communication. The difficulty, therefore, lies in programmatically separating the language used in tweets to convey metaphors from actual news of calamities. In order to determine if a tweet is indeed connected to a crisis, this study presents a machine learning algorithm. The suggested approach uses a painstakingly hand- classified dataset of 10,000 tweets and combines vectorization, classification, and NLP models to improve prediction accuracy. By tackling the challenges presented by metaphorical language, this research helps to construct a sophisticated machine learning framework that can determine the genuine nature of tweets connected to emergencies.



**Keywords:** twitter, real time, text analysis, NLP, disaster detection

# Acknowledgements

An acknowledgements section is optional. You may like to acknowledge the support and help of your supervisor(s), friends, or any other person(s), department(s), institute(s), etc. If you have been provided specific facility from department/school acknowledged so.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| SMPCS | School of Mathematical, Physical and Computational Sciences |

# Chapter 1

# Introduction

Twitter has emerged as a ubiquitous platform for event reporting, facilitated by the widespread use of smartphones. This dynamic environment offers unparalleled opportunities for immediate and decentralized communication, especially during critical events such as disasters. The advent of this digital era has underscored the pressing need for effective crisis communication strategies to harness the potential of Twitter as a valuable tool for situational awareness and emergency response.

However, amidst the wealth of real-time data flowing through Twitter feeds, a significant challenge arises in the accurate identification and differentiation of metaphorical expressions from authentic crisis-related information within tweets. Metaphors, while a powerful linguistic tool for expression, often introduce ambiguity and complexity, posing a considerable hurdle in the quest for reliable crisis detection. The inherent nature of metaphorical language requires a nuanced understanding that transcends traditional analytical approaches, demanding innovative solutions to decipher the true intent behind tweets during critical events.

This research embarks on the journey to address this multifaceted challenge by delving into the intricacies of Twitter communication during crises. The aim is to develop a sophisticated machine learning framework capable of distinguishing between metaphorical language and genuine crisis-related information. Leveraging the prevalence of smartphones and the instantaneous nature of Twitter reporting, this study seeks to contribute to the advancement of crisis communication strategies, fostering a more effective and accurate response to emergencies in the digital age.

## 1.1 Background

The motivation stems from the critical need for effective crisis communication strategies in utilizing Twitter as a valuable tool for situational awareness and emergency response.

The central challenge lies in accurately discerning metaphorical expressions from genuine crisis-related information within tweets. Metaphors, while powerful for expression, introduce ambiguity, posing a significant hurdle to reliable crisis detection. This project addresses this challenge through the development of a sophisticated machine learning framework, drawing on established classification algorithms with intentional omission of specific names for flexibility. Hyperparameter tuning and model selection are explored for optimization. Concurrently, Natural Language Processing (NLP) models capture contextual nuances to enhance metaphorical language understanding.

## 1.2 Problem statement

The significant challenge in the realm of disaster monitoring involves distinguishing disaster-related tweets from general Twitter content. This study aims to develop a machine learning algorithm capable of addressing this challenge and accurately determining if a tweet is genuinely connected to a crisis.

## 1.3 Aims and objectives

**Aims:** The primary aim of this project is to enhance the field of disaster monitoring on Twitter by developing a sophisticated machine learning framework. The goal is to accurately distinguish tweets related to disasters from the broader spectrum of general content on the platform. Through this endeavor, we seek to contribute to the improvement of crisis communication strategies in the digital age.

**Objectives:** Implement data exploration and preprocessing techniques to ensure the dataset is prepared for training and evaluation. Explore and apply various machine learning classification algorithms for the effective categorization of tweets, emphasizing the optimization of hyperparameters and model selection. Incorporate Natural Language Processing (NLP) models to capture contextual nuances, enhancing the understanding of metaphorical language within tweets. Evaluate the performance of the developed framework using rigorous metrics to ensure accuracy and reliability in distinguishing disaster-related tweets. Provide insights and recommendations for advancing crisis communication strategies based on the project outcomes.

## 1.4 Solution approach

The solution approach consists of distinct stages, including data collection, exploration, preprocessing, and the implementation of algorithms for classification and NLP.

### 1.4.1 Data Collection, Exploration, and Preprocessing

**Data Exploration:** Comprehensive exploration is conducted to understand the characteristics of the dataset, including the distribution of metaphorical expressions and crisis-related content. **Preprocessing:** Textual data undergoes preprocessing, including tokenization, stemming, and handling of special characters, to prepare it for subsequent stages.

### 1.4.2 Implementation of Algorithms

Established machine learning classification algorithms are implemented to effectively classify tweets. Hyperparameter tuning and model selection are explored to optimize performance. NLP models are implemented to capture contextual nuances and improve the understanding of metaphorical language in tweets.

## 1.5 Summary of contributions and achievements

This research contributes a sophisticated machine learning framework capable of distinguishing metaphorical language from genuine crisis-related information on Twitter. Achievements include the development of a robust algorithm, leveraging a hand-classified dataset for effective model training.

## 1.6 Organization of the report

The report follows a structured format, exploring background, problem statement, solution approach, detailed methodologies, results, discussions, and conclusions.

# Chapter 2

# Literature Review

Recognizing the imperative for automated solutions to distinguish genuine disaster-related tweets from metaphorical or unrelated content, the research gets crucial reference, aiding in exploring effective approaches and refining the accuracy of disaster-related tweet prediction models from this Kaggle competition Addison Howard (2019) Phil Culliton (2019). With widespread smartphone use, Twitter becomes a primary source for disaster-related information. Utilizing a dataset of 10,000 hand-classified tweets, the goal is to employ machine learning models for binary classification and NLP to predict tweets genuinely related to disasters.

## 2.1 Related Work

In the research Chanda (2021) extensive evaluation of Deep Learning methods for classifying disaster-related tweets. Identification of preprocessing steps, emphasizing named entity substitution.Performance analysis of custom neural networks and Transformer models.Emphasis on practical application potential for automatic disaster detection. BERT is used in Deb and Chanda (2022) where exploration of Twitter's real-time data for disaster identification and challenges in manual data processing due to volume is elaborated. Evaluation of BERT embeddings' superiority in disaster prediction, compared to traditional word embeddings. Discussion on opportunities and challenges of BERT embeddings in Twitter sentiment analysis. Detailed analysis for various algorithms are explored in Fontalis et al. (2023) Theoretical basis of various ML algorithms for tweet analysis (BNB, MNB, LR, KNN, DT, RF).Process flow from dataset import to model training, emphasizing the importance of Exploratory Data Analysis (EDA). Selection of ML models based on suitability, using Wordclouds for identifying relevant words.Explanation of Bayesian algorithms, logistic regression, decision tree, and random forest usage.Iparraguirre-Villanueva et al. (2023)Exploration of BERT embeddings' effectiveness in disaster prediction on social media. Overview of challenges in manual disaster identification due to data volume. Application of different word embeddings (BOW, context-free, contextual) in disaster prediction models. Utilization of embeddings in both traditional ML methods and neural network-based models.In Saddam et al. (2023)Studies introduce normalization processes for words with similar meanings. Stemming, using the Indonesian literary library in Python, maximizes text processing efficieny. Machine learning involves labeling real opinions, crucial for SVM model training. SVM models, especially for multiclass classification, are discussed for sentiment analysis. K- Fold Cross Validation ensures robust testing, evaluating accuracy, precision, recall, and F- score. Confusion matrices aid in a

comprehensive understanding of model performance.

## 2.2 Critique of the review

Existing research demonstrates the effectiveness of advanced techniques such as deep learning models and contextual embeddings for disaster-related sentiment analysis. Critique emphasizes the need for continuous improvement in addressing challenges related to bias mitigation and contextual understanding.

## 2.3 Summary

The literature review underscores the significance of sophisticated preprocessing techniques, advanced text processing, and the application of machine learning models for disaster-related sentiment analysis on Twitter. Key findings highlight the potential of deep learning and contextual embeddings in achieving accurate disaster detection, with practical implications for real-world applications. Continuous improvement is encouraged to address existing challenges and further enhance the reliability of sentiment analysis systems.

# Chapter 3

# Methodology

## 3.1 Working with the Tweets Dataset

### 3.1.1 Explorartory Data Analysis

**Number of Words in Tweets:** To understand the distribution of the number of words in tweets, we plotted histograms for disaster and non-disaster tweets. The histogram for disaster tweets shows a peak around 15-20 words, while non-disaster tweets tend to have a peak around 10-15 words. This indicates that disaster-related tweets may be slightly longer on average compared to non-disaster tweets.

**Punctuations in Tweets:** We analyzed the usage of punctuations in tweets, separating disaster and non-disaster tweets. The histograms displayed the frequency of various punctuation marks such as commas, exclamation marks, and hashtags. Disaster tweets tend to have more question marks and hashtags compared to non-disaster tweets, suggesting a more urgent or expressive tone.

**Common Words in Tweets:** Using a word cloud and bar plot, we identified the most common words in disaster-related tweets. Terms like "fire," "disaster, "shelter" appeared prominently, reflecting the focus on urgent situations and crises in these tweets.

**Bigrams Analysis:** Bigrams analysis revealed common pairs of words occurring together in tweets. The bigrams showed the preposition of the places like "in the", "on the". "to be",indicating discussions about readiness and safety measures. "for the" bigram suggesting discussions about response efforts and support for affected individuals. This analysis provides insights into the topics and contexts discussed in disaster-related content.

Notably, the bigram "http co" appeared frequently, indicating the presence of URLs and links that require cleaning and removal from the tweet text. Handling these URLs is crucial as they often do not contribute to the analysis of tweet content and can introduce noise.

### 3.1.2 Data Cleaning

**Removing URLs** We started by removing URLs from the tweet text as they often don't contribute to the analysis of tweet content and can introduce noise.

**Removing Punctuations** Next, we removed punctuations such as commas, periods, and exclamation marks from the text. This step helps in standardizing the text data and focusing on the actual words and meanings in the tweets.

**Word Cloud after Data Cleaning** After data cleaning, we generated a word cloud to visualize the most common words in disaster-related tweets. This visualization helps in understanding the key themes and topics prevalent in the cleaned tweet data.

These exploratory data analysis and data cleaning steps are crucial for preprocessing the tweet data and gaining insights into the characteristics of disaster-related content on social media platforms.

## 3.2 Algorithms

### 3.2.1 Logistic regression for tweet classification

Logistic regression is a linear classification algorithm used extensively for binary classification tasks, where the goal is to predict a binary outcome (e.g., disaster or non-disaster) based on input features (e.g., tweet text).

**Text Preprocessing and Vectorization:** Text Tokenization:The tweet text is tokenized, breaking it down into individual words or tokens. Count Vectorization: The CountVectorizer is used to convert the tokenized text into numerical vectors. Each tweet is represented by a vector where each element corresponds to the count of a specific word in the tweet.

**Dataset Splitting:** The dataset is split into training and validation sets using a predefined ratio (e.g., 80 percent for training, 20 percent validation). This separation allows us to train the model on a subset of data and evaluate its performance on unseen data.

**Training Process:** The model is trained using the training data and their corresponding vector representations. During training, the model adjusts its coefficients to minimize the logistic loss function and improve classification accuracy.

**Performance Evaluation:** The model's accuracy on the training set is calculated to assess its performance on data it has been trained on. Training accuracy indicates how well the model fits the training data. The model's accuracy on the validation set is computed to evaluate its generalization ability to unseen data. Validation accuracy helps determine if the model can make accurate predictions on new tweets.

**Confusion Matrix Analysis:**

The confusion matrix is generated to visualize the model's predictions on both the training and validation sets. It consists of four quadrants: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

**F1 Score Calculation:**

The F1 score, which combines precision and recall, is calculated based on the confusion matrix. It provides a single metric to assess the model's overall performance, particularly useful for imbalanced datasets.

### 3.2.2 DistilBERT for tweet classification

DistilBERT is a state-of-the-art transformer-based model designed for natural language processing (NLP) tasks, particularly well-suited for tweet classification, such as identifying disaster-related tweets from non-disaster ones.

**DistilBERT Model Configuration** Preprocessing: The DistilBERT model uses a preprocessor specifically tailored for tweet analysis, with a preset configuration for handling tweet sequences up to 160 tokens in length.

**Model Architecture:**

The DistilBERT classifier is based on the "distil bert base en uncased" preset, which utilizes a distilled version of the BERT (Bidirectional Encoder Representations from Transformers) architecture. This architecture incorporates transformer layers to process input tokens, capturing contextual relationships between words and phrases in the tweet. The "distilbert-base-uncased" denotes a base-sized variant of the DistilBERT model that was trained using uncased text data. This model finds widespread application across a range of natural language processing tasks like text classification, sentiment analysis, question answering, among others.

**Training the model:** The training setup includes a batch size of 20, a training split of 80, and a validation split of 20 to ensure robust model evaluation. The number of training examples is determined dynamically from the dataset, and the number of steps per epoch is calculated based on the batch size and training examples. To maintain reproducibility, a random seed of 0 is set before training the model, ensuring consistent results across different runs.

**Model Summary**

The DistilBERT model summary provides insights into the model architecture, including the number of layers, parameters, and output shapes. This summary aids in understanding the complexity and capabilities of the DistilBERT classifier for tweet analysis. The classifier is trained over 7 epochs, optimizing the model using the Adam optimizer with a learning rate of 1e-5. During training, performance metrics such as accuracy and loss are monitored to assess the model's learning progress. Evaluation includes validating the model's performance on a separate validation set to gauge its ability to generalize to unseen data.

## 3.3 Summary

Logistic regression is suitable for simpler text classification tasks, offering interpretability and efficiency, whereas DistilBERT excels in capturing complex patterns and semantics, albeit at the cost of increased computational resources and complexity. The choice between the two depends on the task complexity, dataset size, interpretability requirements, and available computational resources.

The DistilBERT model benefits from this optimization, ensuring that computational resources are utilized effectively during training and inference. Overall, the configuration and training setup of the DistilBERT model for tweet analysis demonstrate a methodical and optimized approach, leveraging state-of-the-art NLP techniques to achieve accurate classification of disaster-related tweets.

# Chapter 4

# Results

The results chapter tells a reader about your findings based on the methodology you have used to solve the investigated problem. For example:

- If your project aims to develop a software/web application, the results may be the developed software/system/performance of the system, etc., obtained using a relevant methodological approach in software engineering.
- If your project aims to implement an algorithm for its analysis, the results may be the performance of the algorithm obtained using a relevant experiment design.
- If your project aims to solve some problems/research questions over a collected dataset, the results may be the findings obtained using the applied tools/algorithms/etc.

Arrange your results and findings in a logical sequence.

## 4.1 A section

...

## 4.2 Example of a Table in LaTeX

Table 4.1 is an example of a table created using the package LaTeX"booktabs." do check the link: wikibooks.org/wiki/LaTeX/Tables for more details. A table should be clean and readable. Unnecessary horizontal lines and vertical lines in tables make them unreadable and messy. The example in Table 4.1 uses a minimum number of liens (only necessary ones). Make sure that the top rule and bottom rule (top and bottom horizontal lines) of a table are present.

Table 4.1: Example of a table in LaTeX

| Bike | | |
|---|---|---|
| Type | Color | Price (£) |
| Electric | black | 700 |
| Hybrid | blue | 500 |
| Road | blue | 300 |
| Mountain | red | 300 |
| Folding | black | 500 |

## 4.3 Example of captions style

- The **caption of a Figure (artwork) goes below** the artwork (Figure/Graphics/illustration). See example artwork in Figure **??**.
- The **caption of a Table goes above** the table. See the example in Table 4.1.
- The **caption of an Algorithm goes above** the algorithm. See the example in Algorithm **??**.
- The **caption of a Listing goes below** the Listing (Code snippet). See example listing in Listing **??**.

## 4.4 Summary

Write a summary of this chapter.

# Chapter 5

# Discussion and Analysis

Depending on the type of project you are doing, this chapter can be merged with "Results" Chapter as " Results and Discussion" as suggested by your supervisor.

In the case of software development and the standalone applications, describe the significance of the obtained results/performance of the system.

## 5.1 A section

Discussion and analysis chapter evaluates and analyses the results. It interprets the obtained results.

## 5.2 Significance of the findings

In this chapter, you should also try to discuss the significance of the results and key findings, in order to enhance the reader's understanding of the investigated problem

## 5.3 Limitations

Discuss the key limitations and potential implications or improvements of the findings.

## 5.4 Summary

Write a summary of this chapter.

# Chapter 6

# Conclusions and Future Work

## 6.1 Conclusions

Typically a conclusions chapter first summarizes the investigated problem and its aims and objectives. It summaries the critical/significant/major findings/results about the aims and objectives that have been obtained by applying the key methods/implementations/experiment set-ups. A conclusions chapter draws a picture/outline of your project's central and the most signification contributions and achievements.

A good conclusions summary could be approximately 300–500 words long, but this is just a recommendation.

A conclusions chapter followed by an abstract is the last things you write in your project report.

## 6.2 Future work

This section should refer to Chapter 4 where the author has reflected their criticality about their own solution. The future work is then sensibly proposed in this section.

**Guidance on writing future work:** While working on a project, you gain experience and learn the potential of your project and its future works. Discuss the future work of the project in technical terms. This has to be based on what has not been yet achieved in comparison to what you had initially planned and what you have learned from the project. Describe to a reader what future work(s) can be started from the things you have completed. This includes identifying what has not been achieved and what could be achieved.

A good future work summary could be approximately 300–500 words long, but this is just a recommendation.

# Chapter 7

# Reflection

Write a short paragraph on the substantial learning experience. This can include your decision-making approach in problem-solving.

**Some hints:** You obviously learned how to use different programming languages, write reports in LaTeXand use other technical tools. In this section, we are more interested in what you thought about the experience. Take some time to think and reflect on your individual project as an experience, rather than just a list of technical skills and knowledge. You may describe things you have learned from the research approach and strategy, the process of identifying and solving a problem, the process research inquiry, and the understanding of the impact of the project on your learning experience and future work.

Also think in terms of:

- what knowledge and skills you have developed
- what challenges you faced, but was not able to overcome
- what you could do this project differently if the same or similar problem would come
- rationalize the divisions from your initial planed aims and objectives.

A good reflective summary could be approximately 300–500 words long, but this is just a recommendation.

**Note:** The next chapter is "**References**," which will be automatically generated if you are using BibTeX referencing method. This template uses BibTeX referencing. Also, note that there is difference between "References" and "Bibliography." The list of "References" strictly only contain the list of articles, paper, and content you have cited (i.e., refereed) in the report. Whereas Bibliography is a list that contains the list of articles, paper, and content you have cited in the report plus the list of articles, paper, and content you have read in order to gain knowledge from. We recommend to use only the list of "References."

# References


Addison Howard, d. (2019), 'Natural language processing with disaster tweets'.
**URL:** *https://kaggle.com/competitions/nlp-getting-started*

Chanda, A. K. (2021), 'Efficacy of bert embeddings on predicting disaster from twitter data', *arXiv preprint arXiv:2108.10698* .

Deb, S. and Chanda, A. K. (2022), 'Comparative analysis of contextual and context-free embeddings in disaster prediction from twitter data', *Machine Learning with Applications* **7**, 100253.

Fontalis, S., Zamichos, A., Tsourma, M., Drosou, A. and Tzovaras, D. (2023), 'A comparative study of deep learning methods for the detection and classification of natural disasters from social media'.

Iparraguirre-Villanueva, O., Melgarejo-Graciano, M., Castro-Leon, G., Olaya-Cotera, S., John, R.-A., Epifanía-Huerta, A., Cabanillas-Carbonell, M. and Zapata-Paulini, J. (2023), 'Classification of tweets related to natural disasters using machine learning algorithms'.

Phil Culliton, Y. G. (2019), 'Natural language processing with disaster tweets'.
**URL:** *https://kaggle.com/competitions/nlp-getting-started*

Saddam, M. A., Dewantara, E. K. and Solichin, A. (2023), 'Sentiment analysis of flood disaster management in jakarta on twitter using support vector machines', *Sinkron: jurnal dan penelitian teknik informatika* **8**(1), 470–479.

# Appendix A

# An Appendix Chapter (Optional)

Some lengthy tables, codes, raw data, length proofs, etc. which are **very important but not essential part** of the project report goes into an Appendix. An appendix is something a reader would consult if he/she needs extra information and a more comprehensive understating of the report. Also, note that you should use one appendix for one idea.

An appendix is optional. If you feel you do not need to include an appendix in your report, avoid including it. Sometime including irrelevant and unnecessary materials in the Appendices may unreasonably increase the total number of pages in your report and distract the reader.

# Appendix B

# An Appendix Chapter (Optional)

...