Master Thesis

Real-time detection of natural disasters from social media using NLP

Faculty of Management, Economics and Sciences
Lille Catholic University

Ahmad Almosa Almaksour & Celina Pruvost

Master Data & AI

2022-2023

Supervised by **Aurelien Vannieuwenhuyze**

Submitted in partial fulfilment of the requirements for the degree of **Master Data & AI** **2022-2023** Supervised by **Aurelien Vannieuwenhuyze**

CERTIFICATE OF NON-PLAGIARISM

I, the undersigned, ALMOSA ALMAKSOUR Ahmad and PRUVOST Celina, students at the FGES, during the academic year "2022/2023", certify that the present Master thesis "Real-time detection of natural disasters from social media using NLP", is strictly the fruit of our personal work, of synthesis and analysis.

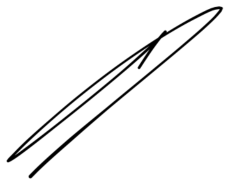
Any quotation (articles, books, memoirs, company documents, Internet sources, etc.) is formally noted as such, explained and referenced in the body of the text and in the bibliography. Any painting or model (photos and various illustrations) is duly cited if it is borrowed from an author or cited as a source if it is adapted.

Any breach of this non-plagiarism Charter will result in the suspension of the evaluation of the thesis, a mark equal to 0, and a summons to appear before the school's disciplinary board.

Done at Lille on 19 April 2023.

Ahmad ALMOSA ALMAKSOUR

Celina PRUVOST



CONTENTS

Abstract

This master thesis explores real-time detection of natural disasters from social media using Natural Language Processing (NLP) techniques. Social media platforms, especially Twitter, provide valuable real-time data that can be utilised to track and monitor disaster situations. The research aims to develop an NLP model to accurately classify tweets related to natural disasters and differentiate them from unrelated or metaphorical content. The study employs various machine learning algorithms such as Naïve Bayes, Random Forest, and Sequential models to classify tweets as disaster-related or not.

The results demonstrate the potential of these models in accurately classifying disaster-related tweets with high precision, recall, and F1-scores. Among the models, the Sequential model shows promising results, effectively distinguishing between different contexts in which the term "wildfire" appears. However, limitations such as small dataset size, imperfect lemmatization, and potential overfitting indicate room for improvement.

Future work could focus on detecting fake news in disaster-related tweets by expanding the training dataset, utilising advanced NLP techniques, incorporating user-related features, and developing an ensemble approach. This would contribute to the development of a comprehensive tool for monitoring and analysing social media content during emergencies, ultimately promoting a safer and more informed society.

INTRODUCTION

Social media has grown in no time over the previous few years and it's still growing significantly, quite half the globe now uses social media which suggests **4.55 billion** people around the world now use social media in keeping with the most recent global social media statistics research summary for 2021 (Chaffey, 2021) [1].

Moreover, social media is one of the foremost important sources of data because it contains a giant mass of knowledge. Due to this, it becomes an invaluable tool for collecting data.

Twitter is one of the most popular microblogging where users create a real-time short message called "Tweet" for sharing opinions about different topics and writing whatever they want, so we can retrieve tweets that contain some information and use it for analysing sentiments using a machine-learning algorithm which can accurately classify tweets as positive or negative (Alec Go, n.d.) [2] .

Therefore, we are able to use social media data for advertising to spot and address specific audiences to see which messages are handiest among certain demographic groups or the best time of day to run a particular ad on a selected digital platform, for building predictive models using sophisticated machine learning algorithms (Ramirez, n.d.) [3]. Moreover, we can use this data more efficiently, maybe for saving lives.

These days, people tend to use social media communication more than other communication tools for telling about dangerous situations like disasters, they pass this information to each other in an extremely short time. As a result, information can be disseminated more rapidly and efficiently during emergencies compared to traditional media outlets such as television or other communication tools (Acar and Muraki, 2011) [4].

We can extract through social media crucial information about natural disasters in real-time, which might save lives thanks to the short time of the message to be received. Social media data allows us also to define some information about disasters like time, location, and type of disaster.

When disaster strikes, social media can provide us with information like images or videos regarding the disaster's time and position of the disaster in real-time. Through Twitter alone, millions of users interact every day, and during disaster times, the amount of interactions and tweets increase to provide information about the situation locations to remain aloof from, where aid is found, locations of high exposures, locations of individuals who need help, and notify emergency personnel of the severity of matters before stepping into that place, it's faster and more efficient than the phone or television (Walter, n.d.)[5] . Social media has been shown to have impact effects during tropical storms around the world. According to Osman Ulvi's (2019) [6] research, it holds the position of the fourth most preferred channel for receiving emergency information in times of crisis or disaster.

Major telecom infrastructure in Pakistan had fallen during the 2014 Kashmir floods. Local officials were unable to communicate with either the government or the army. To contact the government, they used Twitter as a communication medium. Moreover, during the Kashmir floods, the army saved almost 12,000 people based on information from social media. Twitter was also used to coordinate the distribution of humanitarian supplies from throughout the country. For distributing relief items, there were about 40 collection locations across the country and several distribution methods in the flood-ravaged valley. Several lives were saved as a result of the use of social media in this situation (Saleem, n.d.) [7].

Everyone has begun to look into the potential of social media to aid in the aftermath of major natural disasters. People are beginning to consider social media as one of the quickest and easiest ways to convey information amid large-scale disasters. Natural disasters are now shaping new innovations on the social media environment, according to TechRadar (Smith, n.d.) [8]. Twitter's Alert service, Google's Person Finder, and Facebook's Safety Check are all meant to help people in catastrophes alert their family that they are safe and locate others in disasters.

Governments have also been chastised for failing to use social media to distribute disaster alerts and information. After a major landslide hit Sri Lanka in May 2016, the government was chastised for failing to put in place procedures to exchange flood updates, recommendations, and locations. Ordinary citizens had to take up the slack, although it was speculated that the crisis would have been less devastating if government channels had learned how to effectively use social media. Hopefully, things will only get better from here (Thorpe, n.d.) [9].

On the other hand, the trustworthiness of the information being shared is frequently questioned as the mass production of data via social media grows. People share incorrect information on social media for a variety of reasons: some want to grab attention with a spectacular post, others are pursuing a money-making scheme or political objective, and still, others are just repeating faulty or outdated information. Consumers of the postings are led to think in a false world and suspicious conduct as a result of this information. False information can have a variety of characteristics, including doubt in the ”facts,” emotional exploitation of a situation, and so on.

Being first with information is extremely important in today’s media landscape. When crisis rumours begin to circulate, both rookie and experienced users may search the internet, often publishing photographs of the first results without first checking the date or authenticity of the information they are sharing. This is especially common when users share photographs from previous disasters in a hurry as proof of a calamity, which is frequently thought to be factual.

Humanity Road, by Catherine Graham Following the April 2015 earthquake in Nepal, a Facebook post stated that 300 homes in Dhading required assistance. Within 48 hours, the post had been shared over 1,000 times, reaching over 350,000 individuals. The sender of this letter was using social media to seek assistance for locals. According to Facebook data, the typical user has 350 contacts, which means that this single message was seen by roughly 350,000 people. This need had already been published on quakemap.org, a crisis- mapping database established by internet volunteers and administered by Kathmandu Living Labs, a week before the viral post. Helping Hands (a humanitarian organisation) was notified on May 7 and received much-needed basic necessities by May 11. The late Facebook post was intended to be helpful, but the need had already been met. This little example explains how distributing outdated information wastes resources and puts people’s lives at danger (security, n.d.) [10]

STATE OF THE ART

The use of social networks nowadays

According to a study of the *U.S. Bureau of Labor Statistics [11]*, nowadays, people spend as much time on social networks as they do on housework or food preparation. Some use it just to relax and others may use it for a more important cause. In fact, they have become a platform for expression, exchange and information used on a global scale. Indeed, according to the *Digital 2022: Global Overview Report [12]* 4.70 billion people around the world have a social network and this figure has increased by 59% in 5 years from 2.73 billion in 2017 to 4.70 in 2022. People are so connected that they now spend an average of 2 hours 27 minutes a day on social media.

It therefore seems natural to use social networks as a source of information to find out what is happening in the world in real time.

Among the most important social networks in 2022, FaceBook is the leader with 2.91 billion users, followed by Youtube with 2.562 billion users and Whatsapp with 2 billion users. Twitter has 436 million users.

Many of the most used social networks are based on the publication of photographs or videos. However, for the dissemination of textual information in real time, the most relevant networks are Facebook and Twitter. The difference between the two social media is that Facebook allows you to create a homonymous account and connect to your friends/family whereas Twitter allows you to create an anonymous account and connect with unknown people

Although Facebook is the network with the most users, Twitter would also be suitable for detecting natural disasters in real time. The most important thing is the population present on these networks and the type of information published.

All age groups of users are represented on social media, although one stands out the most: the 18-30 year olds. Indeed, based on addressable ad audiences, 55 % of Facebook users and

55.6% of Twitter users are between 18 and 34 years old (*Digital 2021 April Global Statshot Report*) [13].

In addition, about 63-64% of people aged 18-34 use social networks to get news, a figure that tends to decrease with age (*Digital 2022 : July Global Statshot Report*[4])[14]. It is therefore necessary to target the trendy networks to this population. For example, according to *Pew Research Center [15]*, in the US it turns out that in 2021, 55% of users turned on Twitter regularly to get news.

As explained above, Twitter reaches a wider audience of people with a more open circle of followers. All this information gathered shows that the choice of Twitter to study events in real time is very relevant.

Data extraction and preparation

Because the data elements in social media messages are frequently unstructured, extensive analysis is required to extract relevant information. The textual content of a post, for example, may include information about the conversation topic, the sentiment of the user who submitted the post, the location where the post was posted, user opinion on a certain issue, and risk mitigation. In order to collect this data, natural language processing (NLP), neural networks, and deep learning approaches are examples of sophisticated machine learning techniques that should be used.

For identifying meaningful messages from social media streams, this proposed approach To et al. (2017) [16] focuses on the use of matching keywords and hashtags. Furthermore, the authors compare and contrast their proposed matching keywords approach with a learning-based system. A five-step strategy has been developed to evaluate the performance of the two methods: (i) removing spam from the data, (ii) mapping the data to affected and unaffected regions, (iii) filtering irrelevant tweets, (iv) sentiment analysis, and (v) data visualisation. The comparative study considers three types of natural disasters (floods, earthquakes, and wildfires). The authors conclude that the learning-based technique collected a greater number of relevant tweets than the matching-based classification based on the trial findings. On the other hand, the matching-based technique has higher quality when compared to the learning-based approach.

Shekhar and Setty (2015) [17] proposed a system to leverage user's gathered information for damage analysis and estimation, as well as to analyse the attitudes and level of the anguish of system users and persons affected by disasters. Data extraction, sorting, and analysis are the three core modules of the proposed system. Data collection and preprocessing are handled by the data extraction/collection and sorting modules. The third module examines the data in order to extract information about the disaster's scale, dispersion, geotagging, frequency of occurrence, and sentiment evaluation. The authors used a geotagged filter to extract the location of the disaster from the tweets, and K-Nearest Neighbour (KNN) was used for the disaster distribution study.

In the classification of text streams, the choice of informative features is equally crucial. Pekar et al. (2016) [18] gives a detailed examination and comparison of several sorts of textual attributes in the context of natural disaster identification in Twitter text streams. There are five main types of features examined in total: lexical, grammatical, semantic, stylistic, and Twitter information. There are 24,589 tweets in the evaluation data set, with 2,193 of them identified as being from eyewitnesses. The study's classification purpose is to determine whether or not a specific tweet was an eyewitness report. By designing two alternative settings for the experiments, the influence of data heterogeneity is investigated. In the first scenario, the total data set is randomly divided into training and test data in a proportion of 1 to 9 to ensure that both test and training sets contain the same crisis. This also ensures that the distribution of features in test and training is consistent. In the second scenario, any tweets linked to a crisis that were present in the training data are absent from the test set. SVM, Logistic Regression (LR), Random Forest (RF), Naive Bayes, and K-Nearest Neighbour (KNN) are among the five classifiers used to assess performance.

In this article Klein et al. (n.d.) [19] ,Natural Language Processing (NLP) approaches are used to filter Twitter text streams based on the seriousness of the material. The suggested system employs NLP to extract facts and an event identification approach to group posts. The suggested system's primary idea is to group the posts that provide relevant information on the underlying disaster events. To achieve this, the authors use a cutting-edge library called Named Entity Recognizer (NER) which can extract tags referring to person, organisation, and location knowledge to extract emergency information, followed by some preprocessing methods to remove stop words. After that, the matchmaking method is used to assign posts to certain emergency clusters.

Social media contains a giant mass of data and this data can be textual data, numerical data and potentially video or speech data, the major challenge is to collect this data to analyse it and extract useful information from it. This paper Batrinca and Treleaven (n.d.) [20] provides a review of data types that we can find in social media. In general, data subdivides into historic data sets which includes most data stored either manually or automatically such as social financial and economic data, and real-time feeds that is live data feeds from social media streams, news services, financial exchanges, telecommunications services, GPS devices, speech and video. Moreover, this data can be raw data that may contain errors or may be unanalyzed or cleaned data.

The paper presents the principal methods to analyse unstructured textual data such as NLP which focuses on how to design computers to process and evaluate massive volumes of natural language data and is a part of language, computer science, and artificial intelligence that studies how computers and human language interact with each other. They also used the Information Retrieval Process to research word frequency distributions, pattern recognition, annotation and tagging, information extraction, link and connection analysis, visualisation, and predictive analytics.

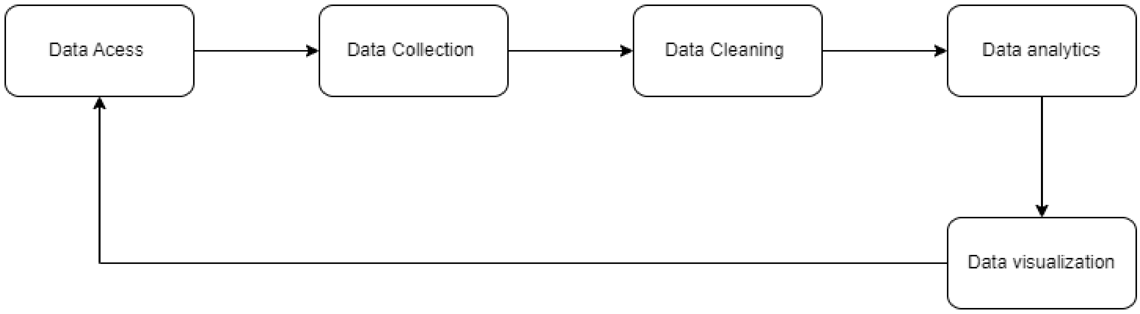


Figure 1 - Social Media Analytics Life Cycle

Moreover, the first step to get data from social media is called *"Scraping"* which is collecting online data from social media and other Web sites in the form of unstructured text. Despite the fact that social media data is available via APIs, the access to this data is a real challenge because of the massive set of data that we find in social media, for example: Twitter has more than 500 million tweets a day which represent a huge amount of information. On the other hand, Twitter recently unveiled the Twitter Data Grants program, where researchers can apply to have access to Twitter's historical data and public tweets in order to obtain new insights from its massive data set.

The next step is cleaning the raw data which is unstructured textual data and the challenge here is to clean data in real time, the paper presents Google Refine and DataWrangler as tools for data cleaning.

This paper [21] represents a framework named *MISNIS - Intelligent Mining of Public Social Networks* which is a platform that mines a large scale of tweets corpus and obtains information and indicators such as user influence or sentiment analysis. This method can retrieve more than 80% of portuguese tweets with a given topic, whereas API restrictions limit access to 1% of all tweets with a given topic. Real-world case studies have shown a 40% increase in the amount of relevant tweets that can be retrieved. In order to collect geolocated tweets, the collection uses Twitter StreamingAPI which allows the access of tweets in real-time, using keywords, hashtags, user ID and geographically delimited region (Kumar 2013) [22]. The API can retrieve only 1% of all tweets but it is potentially possible to retrieve up to 5 million tweets every day using the Streaming API from a single user account, given that there are now 500 million tweets sent every day. Two methods can be used to determine a tweet's geolocation: (1) directly from the tweet when the user chooses to share his location at the time of publishing; and (2) by using the data in the user profile location field. In order to avoid repeated invocations and errors caused by exceeding the API restrictions during the 15-minute window, the built system retrieves timelines using 15 distinct Twitter API access accounts that are synchronised and optimised. When compared to the theoretical 1% freely provided by Twitter, *MISNIS* has been able to collect more than 80% of all streaming Portuguese-language tweets in Portugal through the joint operation of the geolocated data collection and data expansion modules. MISNIS is a work in progress that can be improved in a number of ways: (1) The platform is focused on portuguese language tweets only and it can’t be used with other languages. (2) MISNIS depends on Twitter APIs and any change on these APIs can impact the platform and it modifies and compiles the code.

Marcus et al. (2011) [23] presented *TwitInfo* as an application that enables users to give a list of keywords (such as "obama") and better understand the events associated with those keywords on Twitter. This application is based on another tool *TweeQl* which provides an SQL interface allowing users to send SQL requests with keyword or location and get tweets that match with the given keywords for sentiment analysis.

Identification models using Machine Learning and Deep Learning

In this paper article [24] , the authors tried to detect earthquakes in real time via tweets. First, they removed the stopwords and used stemming to process the sentences. Then, they identified earthquakes by using the two words 'shaking' and 'earthquake'. So if a tweet contains one of the two words it is directly categorised as talking about earthquakes. As a tweet can refer to a past event, users can use these two words in another context. They used a support vector machine to make sure that the author of the tweet really speaks about a current earthquake. They noticed that the shorter the tweet, the more likely it was about a real time earthquake. In addition to the length of the tweet and the position of the keyword, they also looked at the words surrounding the keyword. Researchers also used the location of the tweet and the time of day, this information being sometimes provided.
The results were quite conclusive, however the model only works on the assumption that only one event occurs at a time. It would not be conclusive with a fire for example since two fires can occur at the same time in two different places. Moreover, they used only two target words, the result could be even more relevant if more words on the subject had been targeted.

The goal of Periñán-Pascual & Arcas-Túnez [25] is to detect dangerous situations via tweets in order to help emergency services to act more quickly. Their method is called CASPER (CAtegory- and Sentiment-based Problem FindER) and is based on categorisation but also sentiment analysis. They have evaluated their system on forest fire detection. The CASPER method does not rely on supervised learning.
They identified three different lexicons for hazard identification: the lexicon for each hazard, the lexicon for an emergency situation in general and the lexicon for sentiments. For the emergency lexicon, they built a database from CrisisLex and EMterms and for the feelings

lexicon they used the polarity from SentiWordNet, keeping only unambiguous words. They also added the lexicon of negations that can reverse the meaning of a sentence as well as modifiers that can amplify or minimise a danger. Finally, they took into account abbreviations, which are very common on social networks. The system CASPER consists of 7 steps: First the preprocessing, then the tokenization and the Part Of Speech. In this step, the tweet is in the form of a vector with each word in the tweet and its attributes: place, Part Of Speech, hazard(h), emergency(e) and sentiments(s). The third step is to identify the words concerning the hazard and assign a weight of 1 to the attribute h of each word. The fourth and fifth steps consist of the same as the third step but with the emergency(e) and the sentiments (s). In the sixth step, the impact of the negations and modifiers neighbouring a word is applied to its different attributes.
Finally, a problem relation perception index (PPI) is calculated to measure the degree of reliability and also to set alert thresholds. In the evaluation of tweets about fires, the precision is 0.80.

Traditional machine learning models have therefore already proven their worth in identifying natural disasters through tweets. Deep learning is a subcategory of machine learning that mimics the human learning process through an artificial neural network. It is therefore interesting to see the deep learning approach on our topic of natural disaster identification. This study [26] compared the deep learning approach against a more classical Naive Bayes classification approach in identifying informative tweets during natural disasters. As a training dataset, they used the CrisisLexT26 dataset which contains tweets from 26 disasters that took place in 2012 and 2013. The tweets were annotated according to the type of disaster (natural or not) and whether they were informative or not.
They then trained their model using the Naive Bayes algorithm (this is the classification algorithm that shows the best results). They chose different features: Bag Of Words, Tweet Content Features based on the content of the tweets, User-based Features based on the characteristics of the Twitter account and Polarity-based Features based on sentiment analysis. Regarding the results with Naive Bayes, not all features are equal. The best combination of features used is the Bag Of Words with Polarity and Tweet Content. The average F1-score for informative tweets about natural disasters is 0.902. Note that Bag Of Words alone also scores very well with an average F1-score of 0.900 for the same category of tweets.
They then used Deep Learning models: Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The input of the models is a word2vec model allowing to have a vector

representation of each word. The CNN model showed better results than the RNN model. Compared to Naive Bayes and the selected features, the CNN showed an increase in the F1 score of 3% with an average of 0.928. The deep learning models thus showed a better performance than the classical machine learning models.

Working with a deep learning model seems to be a good first approach given the results of the last study.

Moreover, location of a disaster is crucial information, and tweets can provide valuable data if their point of origin is correctly identified. While a geo-filter tag can be used to extract the location from tweets, it may not always accurately reflect the disaster's location as some tweets may be posted by individuals who are not in the affected region. In such cases, the disaster location may be mentioned within the tweet. This paper [27] proposes a method that utilises the Google Maps Javascript API to identify and label possible locations mentioned in tweets. The proposed method parses each tweet and searches for the coordinates of each word using the Google Maps API. If a word exists in the Google Maps database, the corresponding geographical coordinates are returned, otherwise an error is returned. The identified possible locations are then labelled and associated with their corresponding tweets, allowing for a more accurate analysis of the disaster location.

Gelernter J. and Balaji S. [28] proposed an algorithm for extracting location names from tweets, even down to the level of streets and buildings. The algorithm utilises three distinct modules to identify location mentions: the Named Location Recognizer, which uses lexical pattern matching in a gazetteer trie to identify location names; the Named Entity Recognition (NER) module, which is a natural language processing technique that identifies and classifies named entities such as people, organisations, and locations within text; and rule-based building and street parsers, which use indicator words related to buildings and streets to recognize place names. By combining these modules, Gelernter's algorithm provides detailed and accurate information about the location mentioned in tweets. This innovative approach has numerous applications, including social media analysis and geospatial data mining. Experimental evaluations using Twitter data have demonstrated the effectiveness of the algorithm.

Regulatory requirements

The implementation of a machine learning model is not without precautions, especially when it involves the data of individuals.

Since 2018, the General Data Protection Regulation (GDPR) provides a framework for the processing of personal data within the European Union. The two important terms to define are personal data and data processing.

- A "personal data" is any information relating to an identified or identifiable natural person.
- A "processing of personal data" is an operation or set of operations which relates to personal data, whatever the process used (collection, recording, organisation, storage, adaptation, alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment).[29]

When registering on Twitter, users are obliged to accept the terms of use and the privacy policy. The latter clearly informs users that their tweets are public and freely accessible. They are warned about the use of APIs to make information available to third parties.

When tweets are collected via the Twitter API, it is the collector who then becomes responsible for the data and must therefore comply with the requirements of the GDPR.[30]

The extraction and use of data is subject to a legal framework, and most of the articles discussed above natural disasters from tweets. They collected tweets via an API in which data such as the user's identifier or location, if available, would be stored. This data can be used to identify the author of the tweet and we must therefore comply with the GDPR legislation.

One way to ensure the security of the data that will be retained is through pseudonymisation or anonymisation.

Anonymisation is the removal of personal information that identifies an individual. Once this is done, the data falls outside the legal scope of the GDPR.

Pseudonymisation is the act of assigning an identifier in place of personal data. However,

with the help of additional information, it is possible to trace the data back to the individual. Pseudonymised data still falls within the legal framework of the GDPR.

It would be interesting to retrieve tweets without the username or profile name to have anonymised data as we would like to work also on the location of the tweets.

This state of the art shows that the detection of natural disasters is a current topic that is widely studied. However, to date, we have not found a system that allows us to identify these disasters in real time. The challenges that this entails are the volume of data and identifying words that have a double meaning (metaphor and reality), as well as identifying fake news.

METHODOLOGY

These are the steps of developing a model :

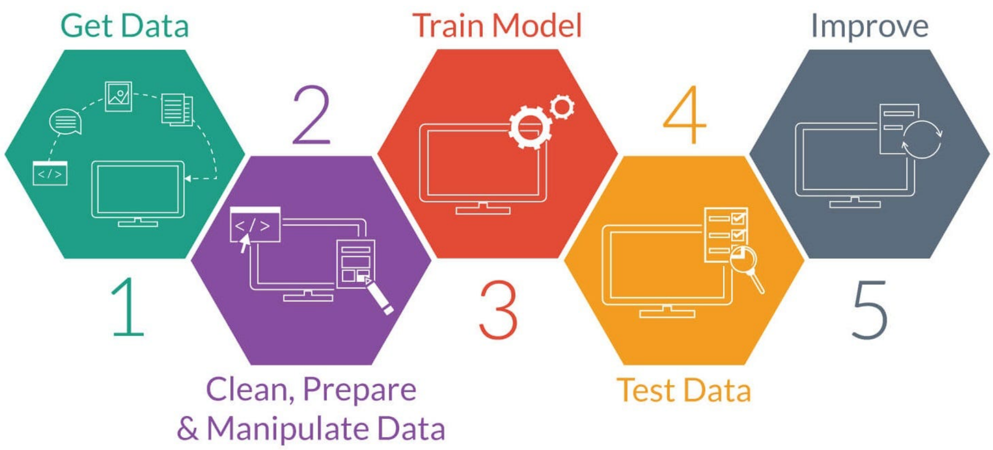


Figure 2 - Steps of developing a ML model

Dataset

In order to train our model, we needed to build up a solid database. We collected tweets via the Twitter API. About 600 tweets were collected. For this, we made a code that retrieves tweets containing the word 'wildfire'. We focused on tweets from the US region using keywords such as 'California', 'New York', 'Texas', 'Florida', 'Colorado'. The tweets were classified as being related to a wildfire. Then, we have recovered tweets that are not linked to a wildfire. These have no topic identity and are therefore in Na. All of these tweets were stored in a dataset. [Appendix 1]

Data preparation

We then had to clean up the data: remove the hashtags, usernames, and web links to keep only the main message of the tweet. We prepared the data so that it would all be processed in the same way using the cleaing_tweets function, thus enabling Natural Language Processing. We first do the tokenization which splits the tweets into smaller units called tokens and then all the cleaning is applied. The output of this function is :

- Tweets without stopwords : the most common words used in the English language
- Tweets without capital letters
- Words of more than two characters or less than twenty characters
- Lemmatized words : this allows us to have the root of the words and thus to limit the number of possible entries for words with the same meaning.

We mainly used the NLTK library which is a library for working specifically with human language data.

Models

We tried 3 different models from the simplest to the most complex:

- Naive method: create a word list containing words in the wildfire lexical field and if a word in the tweet matches, then a wildfire will be identified.

- Machine Learning: after cleaning our data, we transformed it into a Bag of Words or TF-IDF.

The Bag of words method is a way of representing text data as a collection of tokens. It results in a vector that represents the frequency of each word in the document.

The TF-IDF method allows us to take into account all the words, like a Bag Of Words, but giving them a different importance according to their frequency in the set of documents, for us the tweets.

The learning on our data was done in a supervised way, they were labelled. We divided our dataset into train/test with a respective ratio of 80/20.

Then, we tested different machine learning algorithms: Random Forest, Naïve Bayes.

Naïve-Bayes : The determination of the probability of an event occurring based on an event that has already occurred. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Random Forests : it constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees.

- Deep Learning:

In this study, we developed a deep learning model to classify text data using a Sequential model from the Keras library. The model's architecture consisted of the following layers:

- An Embedding layer with input dimensions equal to the vocabulary size, output dimensions as the embedding dimensions, and input length as the maximum length of input sequences.
- A GlobalMaxPool1D layer for reducing feature dimensions.
- Two Dropout layers with 0.4 dropout rates to prevent the overfitting.
- Two Dense layers having 10 neurons with ReLU activation and the number of classes as output neurons with softmax activation, respectively.

We compiled the model using the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric. We then trained the model on the padded training data and corresponding categorical training labels, using a batch size of 50 and 100 epochs, with validation data provided for monitoring performance.

The different models are trained and compared according to their metrics. Once we get a good model and in order not to lose it, it is downloaded as a pickle. It enables users to serialise, i.e. converting into stream of bytes and deserialize Python objects, making it easy to save and restore data. Thus we will only have to load the model to find it again at any time. To keep the same conditions as in the model training, the vectorization must also be downloaded to maintain the same features. [Appendix 2]

Data evaluation

We evaluated our machine learning's results with a confusion matrix and a classification report including accuracy, precision, recall and F1-score. The differences between these scores are :

- Accuracy is the proportion of correct predictions made by the model.
- Precision is the proportion of positive identifications that were actually correct.
- Recall is the proportion of true positives that were correctly identified.
- F1-score takes into account Precision and Recall in its calculation. It is to be used when categories are unevenly distributed and if a balance between precision and recall is sought.

We evaluated the sequential model's performance on the padded test data and categorical test labels, obtaining test loss and accuracy scores.

To evaluate under real conditions, we also tested our different models on sentences to see how they performed.

Location detection

The main goal of our algorithm is to detect wildfires in real time, which is important in order to take immediate action such as dispatching emergency services or evacuating the affected population to minimise the impact.

Once we have identified tweets that are talking about wildfires, we use another function called the "location" function to determine the location of the fire. This function takes the original tweet as input, without any cleaning or preprocessing, and tokenizes the tweet. The tokens are then labelled using the Stanford library's Name Entity Recognizer tool, which can identify whether a token represents a person, location, or organisation ...

Tokens labelled as "location" are then stored and a process is applied to combine the tokens of locations mentioned in two words, such as "Los Angeles". This enables us to identify the specific location of the fire, which is then added as a new column called "location." This information can be used to quickly and efficiently respond to the wildfire and minimise its impact on the affected area.

All of these steps, from data preparation to location detection were made on the same notebook.[Appendix 3]

RESULTS

Overview results

We started by cleaning our tweets from all the elements like #, @ ... and then we did the word processing. To give you an idea of the transformation :

original tweet	cleaned tweet
Northeast Pa wildfire that closed turnpike has burned 2513 acres and was 10 contained	northeast wildfire closed turnpike burned 2513 acre contained

Tweets are divided into two classes : about a wildfire (1) and not (0). Here is the distribution of the classes: 304 tweets are about a wildfire and 268 are not. The classes are therefore equally balanced, i.e. there are approximately the same number of tweets in each class.

```
1.0    304
0.0    268
Name: topic, dtype: int64
```

Figure 3 - Distribution of classes

Here are the most used words in the dataset:

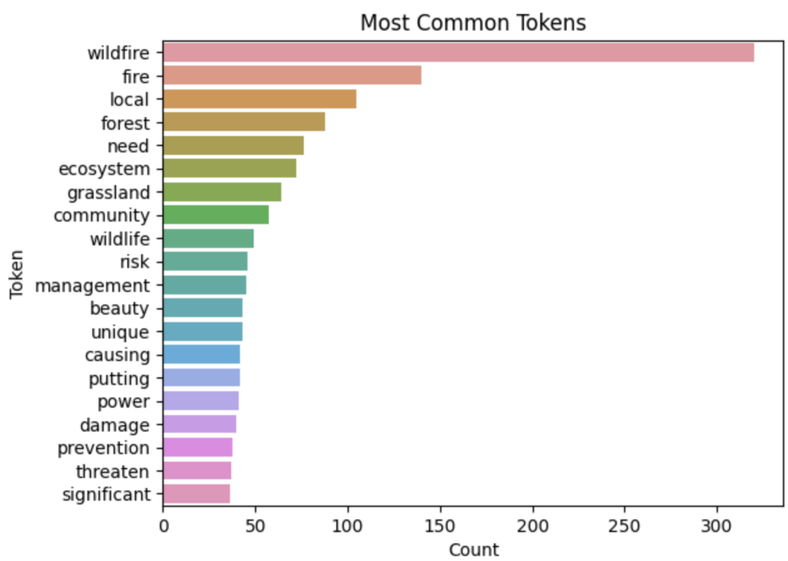


Figure 4 - Most Common words in the dataset

Of the 572 tweets, once the tweets have been cleaned, there are 1374 tokens left. We notice that the word 'wildfire' is present 320 times and the word 'fire' 160 times. This means that the words directly meaning fire represent 480 tokens, i.e. 35% of the tokens.

Results of the Naïve method

Not surprisingly, this method is not very effective.

	precision	recall	f1-score	support
0.0	0.95	0.39	0.56	268
1.0	0.65	0.98	0.78	304
accuracy			0.71	572
macro avg	0.80	0.69	0.67	572
weighted avg	0.79	0.71	0.68	572

```
[[105 163]
 [  5 299]]
```

Figure 5 - Evaluation metrics of Naive method

We can note that this method works better for identifying "wildfire" as such than for identifying other tweets. They tend to be classified as "wildfire" quite easily. Indeed, almost 40% of tweets "other" were classified as "wildfire".

Results of the Naïve Bayes model

This method was based on a bag of words.

Accuracy: 0.9565217391304348

	precision	recall	f1-score	support
0.0	0.94	0.96	0.95	51
1.0	0.97	0.95	0.96	64
accuracy			0.96	115
macro avg	0.96	0.96	0.96	115
weighted avg	0.96	0.96	0.96	115

Figure 6 - The evaluation metrics of Naïve Bayes Model

By looking at the performance evaluation metrics of our model, the accuracy, precision, recall and f1-score are all high (around 0.96).

A high accuracy score suggests that the model is performing well in terms of correctly predicting the classes of the test data. A high recall score indicates that the model is able to correctly identify the class “1” more often than not. A high precision indicates that when it predicts a class “1”, it is usually correct.

Overall, a model with high accuracy, recall, and high precision is considered to be a good performing model.

Results of the Random Forest model

```
Accuracy: 0.9652173913043478

              precision    recall  f1-score   support

         0.0       0.98      0.94      0.96        54
         1.0       0.95      0.98      0.97        61

    accuracy                           0.97       115
   macro avg       0.97      0.96      0.97       115
weighted avg       0.97      0.97      0.97       115
```

Figure 7 - The evaluation metrics of Random Forest model

Like the model using the Naïve Bayes method, the model using Random forest also shows very good results which are even better since the different metrics reach an average of 0.97.

Results of Sequential model

The training outcomes revealed that the model attained a remarkable validation accuracy of 94.19% by epoch 25, which did not see significant improvements in subsequent epochs. Furthermore, based on the figure 8 representation, there seems to be a divergence between the training and validation losses. The training loss continues to decrease, while the validation loss reaches a plateau, starting around epoch 25. This indicates the potential for overfitting, as the validation accuracy plateaus while the training accuracy keeps progressing, achieving 100% at epoch 32. To address this issue, one can employ early stopping or incorporate additional regularisation techniques to prevent overfitting and enhance the model's adaptability to new unseen data.

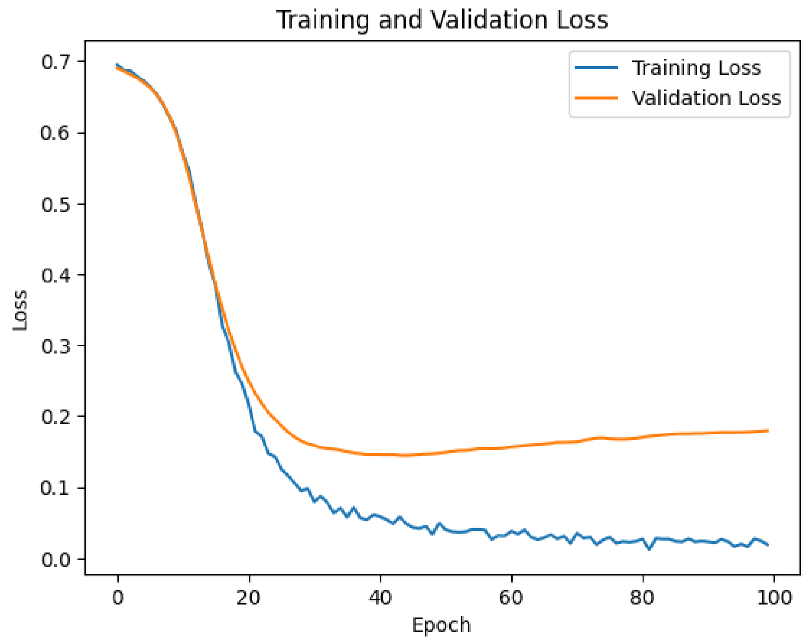


Figure 8 - The training and validation loss of Sequential model

Tests on a sentence

Even if the different models perform well in terms of metrics, it is important to ensure that they perform well by testing them. To do this, we took 3 sentences to try:

- [1] one sentence that is relative to a wildfire *"I just saw a huge wildfire in the forest near my house!"*
- [2] a sentence that is not related to a wildfire *"I really enjoy walking in the forest"*
- [3] a more tricky sentence that contains the word wildfire but is not related to a wildfire *"When the rumour that the company was going to lay off some of its staff spread like wildfire, the employees started to panic"*

Here are the results of the different models :

	Sentence [1]	Sentence [2]	Sentence [3]
Naive method	1	0	1
Naive-Bayes	1	1	0
Random Forest	1	0	0
Sequential	1	0	0

The naive method is the fastest to build but the least accurate. As it only looks for the presence of certain words without taking into account the context, the sentence as a whole, it will lead to many false positives. As soon as there is the word 'fire', it will be interpreted as a real fire whereas this word can be used in several contexts like 'i'm on fire'. Moreover, even if it identifies a fire thanks to fire, this does not prove that it is a forest fire, it could be a domestic fire.

The results are better with Naive-Bayes method but there are still some errors that are surprising. While we might have thought that the model would have been wrong on the tricky sentence, it classified it correctly but misclassified the second one. It seems that the word 'forest' fooled the model and it identified it as an important feature to classify the sentence as 'wildfire'.

The results are even better with the Random Forest. Beyond the fact that the metrics were even better than Naive Bayes, the 3-sentence test was passed and the model classified them all well. To understand how the model works, we looked at its features' importance. The more important a word is, the more impact it will have in the classification.

Here are the 20 most important words in our Random Forest :

```
['devastating',
 'explored',
 'watched',
 'grassland',
 'damage',
 'evacuation',
 'beauty',
 'power',
 'risk',
 'wildlife',
 'ecosystem',
 'causing',
 'threatens',
 'concern',
 'threaten',
 'forest',
 'need',
 'local',
 'fire',
 'wildfire']
```

Figure 9 - The 20 most important tokens for Random Forest model

Not surprisingly, the most important words according to the model to classify whether the sentence is about a wildfire or not are the words: “fire”, “wildfire”, “forests”, “threaten” or “evacuation”. Based on these results, we can test an even trickier sentence with these words and see how the model performs. Note that these words are very similar to the top most common tokens seen above.

As this is the most promising model at this stage, we decided to take a closer look at it.

After multiple trials on different more or less tricky wildfire sentences, we realised that the model was more likely to put in 0 than in 1. It appeared that for the model to classify as "wildfire", the words "wildfire" alone is not enough. For example, the same sentence as the previous one "*I just saw a huge wildfire near my house!*" without the "forest" part is classified as 0 while it is unequivocal about a wildfire. Even if the model seemed to perform well, it is not perfect and many classification errors will be made.

The results presented of the Sequential model demonstrate the capability of the developed model to accurately classify tweets related to real wildfires and differentiate them from those that are unrelated or contain metaphorical uses of the term "wildfire." The model was able to correctly identify a tweet explicitly mentioning a wildfire, a tweet discussing a forest without reference to wildfires, and a tweet using "wildfire" metaphorically.

The model's performance on the three tweets can be attributed to its ability to capture the semantic information and context within the text data through the use of various layers, such as embedding, convolutional, and dense layers. By extracting meaningful features from the raw text and understanding the patterns within the data, the model can make accurate predictions about the context of the term "wildfire" in each tweet.

These findings suggest that the model is effective in distinguishing between various contexts in which the term "wildfire" appears. The successful classification of the provided examples indicates that the model has learned relevant features from the training data and is capable of generalising its knowledge to previously unseen instances. Consequently, this model can be considered a valuable tool for monitoring and analysing social media content related to wildfires.

Location

To evaluate the location detection capabilities of our integrated natural language processing models, we conducted experiments on three sentences that were previously used for testing the performance of different models. A slight modification was made to these sentences as they initially did not contain any location information. For instance, the first sentence was identified as pertaining to a "wildfire" by all the models, so we replaced "my house" with "Los Angeles" to better assess the location detection functionality in our test.

The predictions in our experiments were generated using the Random Forest model.

	sentence	cleaned_sentence	prediction	location
0	When the rumour that the company was going to ...	rumour company going lay staff spread like wil...	0.0	
1	I really enjoy walking in the forest	really enjoy walking forest	0.0	
2	I just saw a huge wildfire in the forest near ...	saw huge wildfire forest near los angeles	1.0	[Los Angeles]

Figure 10 - Identification of location from disaster-related tweet

We can see that the function worked correctly, it was not applied to the first two rows since they are not tweets about wildfires, and on the third row it successfully identified "Los Angeles".

The findings from these experiments not only demonstrated the feasibility of incorporating location detection into our model but also highlighted the importance of geographical context in understanding and responding to disaster events. As a result, the integration of location information adds significant value to our model, making it a more comprehensive and practical tool for disaster detection.

DISCUSSION

In comparison with other studies referenced in the state of art, we developed a sequential deep learning model that performed remarkably well on our dataset, despite the presence of overfitting. As mentioned in [26], deep learning algorithms have demonstrated significant improvement compared to traditional machine learning models, although this generalisation can vary depending on factors such as the specific dataset and other conditions. The model we implemented supports this theory, affirming the potential of deep learning in disaster-related tweet analysis. The implementation of a sequential deep learning model aligns with the findings of Neppalli et al. [26], who found that Deep Neural Networks (DNNs) outperformed Naïve Bayes classifiers when identifying informative tweets during disasters.

However, our work adds to the growing body of evidence supporting the use of deep learning algorithms in disaster-related tweet analysis. By building on the findings of previous studies, such as [26], we have demonstrated the potential of a sequential deep learning model in identifying informative tweets during disasters, even in the presence of overfitting. This outcome encourages further exploration of advanced techniques and strategies to enhance disaster prediction and response capabilities using social media data. Our primary objective was to develop a real-time disaster detection system using tweets. However, due to the challenges associated with obtaining data from the Twitter API and time constraints, we were unable to fully achieve this goal within the scope of this study.

Dataset

Firstly, for optimal training, we should have collected many more tweets: 10,000 would have been ideal. However, the Twitter API has been updated recently and the academic search version that allowed us to retrieve a large number of tweets no longer seems to exist. We both had Twitter Developer accounts but only one of the two authors was able to collect tweets, the other had this error: ”You currently have Essential access which includes access to Twitter API v2 endpoints only. If you need access to this endpoint, you'll need to apply for Elevated access via the Developer Portal”, and this is while both trying to create new accounts.

We assume that the author who was able to collect the tweets was on an older version, as the account was created a long time ago. In addition, the number of tweets that could be collected was limited, which is why we had a small dataset. We therefore trained our models on only 572 tweets, which is a small amount of data and this can lead to several problems, such as overfitting.

Overfitting occurs when a model learns to fit the training data too closely which leads to poor performance on new data. Overmore, this can lead to the model not seeing enough variety in the data and therefore fails to generalise to new, unseen data. Especially as our model is based on NLP, language is a complex thing and beyond the fact that the model will only have encountered a small number of words, it is also necessary to take into consideration the relationships between words and the different meanings that a same word can have; something that will not be well done on a small dataset.

As we were not able to retrieve many tweets, all those collected were saved in our dataset. If we had more tweets, we could have had a validation dataset. Like our tests on sentences, the aim is to provide an unbiased estimate of the model's performance on new and unseen data but on a larger scale. It evaluates the model on data that is not used for training or testing and serves as a checkpoint to help identify potential overfitting issues. It can also be used to tune the model's hyperparameters to improve its performance.

All this could have prevented the overfitting that we see in the different models and therefore we could have had better performing models

Cleaning

When you look at the cleaning, you realise that the lemmatization is not perfect. Indeed, even after applying the function, words_cleaning(), for example, are still present. As mentioned above, this method is supposed to return the root of a word: "run," "running," and "ran," should be returned in a single 'run' form. However, just by looking at the 20 most present tokens, we find the word 'putting'. So the models were trained with many words that had the same root. The purpose of lemmatization is to help improve the accuracy of algorithms by reducing the number of unique words that need to be analysed. We used the

WordNetLemmatizer from NLTK. There are many other packages for lemmatization: TextBlob, Spacy, StanfordCoreNLP. Perhaps we should try our model again using another lemmatization package.

Fake news

While the current study focused on predicting disasters from tweets, an essential avenue for future work would be to extend the model to detect fake news as well. In the era of digital communication, fake news has become a pervasive issue, with individuals spreading rumours on the internet for various reasons. These unfounded claims can have severe consequences, impacting people's lives and causing unwarranted panic or misinformation.

Detecting the authenticity of disaster-related tweets is particularly crucial, as it can enable authorities and emergency responders to concentrate their efforts on addressing genuine incidents and ensuring public safety. Moreover, distinguishing between real and fake disaster news can help prevent the spread of misinformation and contribute to a more informed and prepared society.

To achieve this, the model could be enhanced by incorporating additional features and techniques that capture the subtleties of language and context, which may help differentiate between genuine disaster reports and fabricated news. Some potential approaches include:

- Expanding the training dataset: Including a more diverse and balanced dataset containing both genuine disaster-related tweets and fake news can help improve the model's ability to recognize and discern between the two types of content.

- Utilising advanced natural language processing techniques: Leveraging state-of-the-art techniques, such as transformers and pre-trained language models like BERT or GPT, can enable the model to capture more nuanced contextual information and achieve higher accuracy in detecting fake news.

- Incorporating user-related features: Analysing the credibility and history of the users posting the tweets can provide valuable insights into the likelihood of the content being genuine or fabricated.

- Developing an ensemble approach: Combining the predictions of multiple models or using different classifiers to capture various aspects of the problem can result in a more robust and accurate fake news detection system.

By extending the current model to detect fake news in disaster-related tweets, future work can contribute to the development of a comprehensive and effective tool for monitoring and analysing social media content in emergency situations, ultimately promoting a safer and more informed society.

Computer vision

To further enhance our approach, we can incorporate computer vision techniques, allowing us to detect natural disasters based on images provided within tweets. This leads to more accurate predictions and precise identification of disaster locations, which is crucial for local authorities to respond promptly. Typically, when people tweet about disasters, they do not mention the exact location or the severity of the situation. For instance, someone might tweet about a fire in Lille, but without specifying the location within the city or the urgency of the situation, it becomes challenging to direct assistance to the appropriate area and alert the affected individuals.

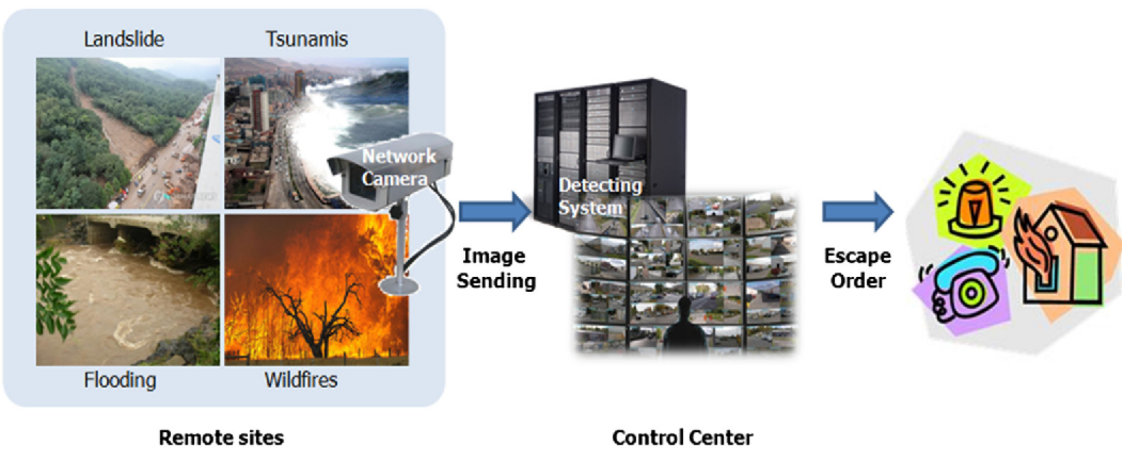


Figure 11 - An example of a detecting system based on computer vision [31]

Using computer vision techniques allows us to handle several vital elements related to disaster detection:

- Disaster classification: By analysing the images associated with a disaster, computer vision can accurately identify and categorise the type of disaster, such as flood, wildfire, earthquake, or hurricane. This information is crucial for authorities and emergency responders to deploy appropriate resources and strategies.

- Severity assessment: Computer vision can analyse visual cues in the images to determine the severity of a disaster. For instance, in the case of flooding, the model can estimate the water depth, extent of inundation, and potential damage to infrastructure. This helps prioritise response efforts and allocate resources effectively based on the urgency of the situation.

- Location detection: Geolocating the disaster is essential for coordinating rescue operations and evacuation plans. Computer vision can extract geographical information from images, such as landmarks, terrain features, or GPS metadata, to pinpoint the disaster's location with high precision.

Integrating natural language processing (NLP) with computer vision offers a powerful approach to disaster management. NLP can process and analyse textual information from social media. When combined with the visual insights from computer vision, we can create a sophisticated model that delivers a comprehensive understanding of the disaster, its precise location, and the level of severity.

However, implementing such a model is not without its challenges, as it requires significant effort and is inherently more complex. Nonetheless, the potential benefits of integrating NLP and computer vision make it a worthwhile pursuit for enhancing disaster prediction and response capabilities.

Tweets geolocalisation

We have not tested our location function on many tweets and it deserves to be further investigated. We do not know its effectiveness on small towns or on abbreviations (LA instead of Los Angeles). Furthermore, in some cases, people might simply not mention their location in the tweet, which makes it difficult to determine the exact location of the fire. In such cases, the coordinates provided by the tweet can be helpful. When users turn on the location functionality while tweeting, the coordinates of their location are attached to the tweet. This information is stored in the status object.

To make use of this information, we can extract the coordinates from the tweets that are geotagged and convert them into a location. The Google Maps API, for example, can convert geographic coordinates into a human-readable address.

Once we have the human-readable address, we can add it to the "location" column along with the locations identified by the Stanford NER Tagger. These two methods are complementary and this will provide a more precise location of the fire. In summary, the use of geotagged tweets can be a valuable addition to our algorithm in identifying the location of the fire.

Sentiments analysis

Incorporating sentiment analysis into the disaster prediction and fake tweet identification model is crucial for understanding people's emotions and reactions to events, as they often reflect the reality of an ongoing disaster. Analysing the sentiments expressed in tweets can provide valuable insights into public response to events and help differentiate between genuine disaster-related content and fake news.

However, it is essential to consider that sentiments expressed on social media can sometimes be biased, as emotions may not always align with the severity of the situation. For instance, in cases where two non-allied countries are involved, some individuals might express negative emotions or even make light of a disaster affecting the other country. Although such behaviour can be disheartening, it is necessary to account for these possibilities when analysing sentiment data.

To integrate sentiment analysis into the disaster detection model, we can leverage pre-trained sentiment analysis models or advanced transformer-based models like BERT or GPT, fine-tuned for sentiment classification tasks. By incorporating sentiment analysis, we can improve the model's ability to differentiate between genuine disaster-related tweets and irrelevant or unrelated content, as well as detect potential biases in public response.

To address the potential biases in sentiment data, we can consider incorporating additional features or techniques that account for the cultural and political context of the tweets. For example, we could analyse the geographical distribution of sentiments or apply unsupervised clustering algorithms to identify different sentiment patterns within the data. These approaches could help reveal biases or anomalies in public response that might not be immediately evident from sentiment analysis alone.

CONCLUSION

This master thesis aimed to develop a real-time natural disaster detection system using social media data, particularly focusing on Twitter as a source of information. The study employed natural language processing (NLP) techniques to analyse and classify tweets to identify whether they pertain to natural disasters, such as wildfires. The significance of this research lies in the potential of such a system to save lives by providing timely and accurate information about ongoing disasters, allowing for faster and more effective responses from authorities and emergency services.

Various machine learning models, including Naïve Bayes, Random Forest, and Sequential model, were applied to analyse the tweets and identify the most suitable approach for this task. The results demonstrated the effectiveness of these models in classifying disaster-related tweets, with the Sequential model outperforming the others in terms of accuracy, recall, and precision. This model's ability to capture semantic information and context within the text data through the use of embedding, convolutional, and dense layers enabled it to distinguish between various contexts in which the term "wildfire" appears, making it a valuable tool for monitoring and analysing social media content related to wildfires.

However, this study also faced several challenges and limitations, including a small dataset, overfitting issues, and imperfect lemmatization. Despite these challenges, the thesis serves as an essential foundation for future work in this area. In particular, it paves the way for the development of a comprehensive natural disaster detection system that also addresses the issue of fake news detection, thereby preventing the spread of misinformation and promoting a safer and more informed society.

By expanding the training dataset, using advanced NLP techniques, incorporating user-related features, and developing an ensemble approach, future research can build upon this thesis to create a more robust and accurate disaster detection system. Ultimately, the findings of this study underscore the potential of harnessing the power of social media and NLP to enhance disaster response and preparedness, contributing to a world where people are better equipped to face and overcome the challenges posed by natural disasters.

REFERENCES

[1] Chaffey, D. (2021). Global social media statistics research summary 2022.
URL: *https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/*

[2] Alec Go, Lei Huang, R. B. (n.d.). Twitter sentiment analysis.
URL: *https://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf*

[3] Ramirez, R. (n.d.). Social media data mining: Understanding what it is and how busi- nesses can use it. **URL:** *https://www.sandiego.edu/blogs/business/de*

[4] Acar, A. and Muraki, Y. (2011). Twitter for crisis communication: lessons learned from japan's tsunami disaster, *Int. J. Web Based Communities* **7**(3): 392–402.
URL: *https://doi.org/10.1504/IJWBC.2011.041206*

[5] Walter, L. (n.d.). Red cross: Web users increasingly rely on social media to seek help in a disaster.
URL: *https://www.ehstoday.com/emergency-management/article/21904762/red-cross-web-users-increasingly-rely-on-social-media-to-seek-help-in-a-disaster*

[6] Osman Ulvi, N. L. (2019). The role of social and mainstream media during storms, *Int. J. Web Based Communities* .
URL: *https://jphe.amegroups.com/article/view/5543/htmlB13*

[7] Saleem, V. (n.d.). How people turn to social media during natural disasters.
URL: *https://www.downtoearth.org.in/blog/how-people-turn-to-social-media-during-natural-disasters-49587*

[8] Smith, C. (n.d.). Dealing with disaster: how social media is helping save the world.
URL: *https://www.techradar.com/news/internet/dealing-with-disaster-how-social-media-is-helping-save-the-world-1203809*

[9] Thorpe, J. (n.d.). 5 ways social media has saved lives.
URL: *https://www.bustle.com/articles/175637-5-ways-social-media-has-saved-actual- lives*

[10] security, H. (n.d.). Countering false information on social media in disasters and emer- gencies.
URL: *https://www.dhs.gov/sites/default/files/publications/SMWG*$_C$*ountering* $-$ *False* $-$ *Info* $-$ *Social* $-$ *Media* $-$ *Disasters* $-$ *Emergencies*$_M$ *ar*2018 $-$ 508*.pdf*

[11] Average hours per day spent in primary activities for the civilian population, 2021 quarterly and annual averages. **URL** :https://www.bls.gov/news.release/atus.t12.htm Accessed on 19 November 2022.

[12] Digital 2022: Global Overview Report
URL : https://datareportal.com/reports/digital-2022-global-overview-report
Accessed on 20 November 2022

[13] Digital 2021: April Global Statshot Report
URL: https://datareportal.com/reports/digital-2021-april-global-statshot
Accessed on 20 November 2022

[14] Digital 2022: July Global Statshot Report
UR: Lhttps://datareportal.com/reports/digital-2022-july-global-statshot
Accessed on 20 November 2022

[15] Pew Research Center, September, 2021, “News Consumption Across Social Media in 2021”

[16] To, H., Agrawal, S., Kim, S. H. and Shahabi, C. (2017). On identifying disaster-related tweets: Matching-based or learning-based?, *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pp. 330–337.

[17] Shekhar, H. and Setty, S. (2015). Disaster analysis through tweets, *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1719–1723.

[18] Pekar, V., Binner, J. and Najafi, H. (2016). Detecting mass emergency events on social media: One classification problem or many?

[19] Klein, B., Castanedo, F., Elejalde, I., L´opezde Ipin˜a, D. and Prada Nespral, A. (n.d.). Emergency event detection in twitter streams based on natural language processing, Lecture Notes in Computer Science.

[20] Batrinca, B. and Treleaven, P. C. (n.d.). Social media analytics: a survey of techniques, tools and platforms, **30**.
URL: *https://doi.org/10.1007/s00146-014-0549-4*

[21] Joao P. Carvalho, Hugo Rosa, Gaspar Brogueira, Fernando Batista, MISNIS: An intelligent platform for twitter topic mining, Volume 89, 2017, Pages 374-388, https://doi.org/10.1016/j.eswa.2017.08.001.

[22] Kumar, S., Morstatter, F., and Liu, H. (2013a). Twitter Data Analytics. Springer, New York, NY, USA.

[23] Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., and Miller, R. C. (2011a).

Tweets as data: Demonstration of tweeql and twitinfo. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD ’11, pages 1259– 1262, New York, NY, USA. ACM

[24] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. The Web Conference. https://doi.org/10.1145/1772690.1772777

[25] Periñán-Pascual, C., & Arcas-Túnez, F. (2018). The Analysis of Tweets to Detect Natural Hazards. Intelligent Environments. doi:10.3233/978-1-61499-874-7-87

[26] Neppalli, V.K., Caragea, C., & Caragea, D. (2018). Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters. International Conference on Information Systems for Crisis Response and Management.

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

[28] Gelernter J. and Balaji S. "An algorithm for local geoparsing of microtext." GeoInformatica : 1-33, 2013.

[29] Art 5. From General Data Protection Regulation

[30] Jonasson, F. (2019). A system for GDPR-compliant collection of social media data: from legal to software requirements.

[31] Ko, ByoungChul and Sooyeong Kwak. “Survey of computer vision–based natural disaster warning systems.” Optical Engineering 51 (2012): 070901-1-070901-11.

APPENDIX

Link to the general project :

https://drive.google.com/drive/u/0/folders/1vps5xnkie3LKUVUKeiEM_KUiWfhKcOB8

[Appendix 1] Link to the dataset :

https://drive.google.com/file/d/1d3WPDGhI3wSwsTcGS7dhHzEeJaamnbDv/view?usp=share_link

[Appendix 2] Links to the pickles:

Bag-Of-Words :

https://drive.google.com/file/d/1RMhacRhJHCHikXkFFL6S2V883OPz-9g7/view?usp=share_link

Naïve bayes:

https://drive.google.com/file/d/14Q5sD1HkBIiDb44dAv-BkgUhsB8ZvZkA/view?usp=share_link

TF-IDF:

https://drive.google.com/file/d/1S26EsV17pVy4S7abOeClgfqai8JCH8AR/view?usp=share_link

Random Forest :

https://drive.google.com/file/d/1Jm6CkubTrs0uHeA8-sYezhylj3KK8clH/view?usp=share_link

[Appendix 3] Link to the main notebook

https://colab.research.google.com/drive/1vQ-o0gh2FyiVBy1yyrqMd2CEjtRASR4v?usp=share_link