Disaster Detection System using Social Media, Machine learning and Crowdsourcing

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

Ajesh M (38119001)
Akhil Anand (38110018)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

MARCH – 2021

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with “A” grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Ajesh M (38119001)** and **Akhil Anand (38110018)** who carried out the project entitled “**Disaster Detection System using Social Media, Machine Learning and Crowdsourcing”** under my supervision from November 2021 to March 2022.

Internal Guide

Dr. T. Prem Jacob M.E., Ph.D.

Head of the Department

Dr. S. Vigneshwari M.E., Ph.D., and Dr. L. Lakshmanan M.E., Ph.D.,

---

Submitted for Viva voce Examination held on ____________________

Internal Examiner **External Examiner**

DECLARATION

I **Ajesh M (Reg No:38119001) and Akhil Anand (Reg No: 38110018)** hereby declare that the Project Report entitled **“Disaster Detection System using Social Media, Machine Learning and Crowdsourcing”** done by us under the guidance of **Dr. T. Prem Jacob M.E., Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in 2018-2022.

DATE:

PLACE: **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to the **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E.,Ph.D.**, **Dean**, School of Computing **Dr.S.Vigneshwari M.E., Ph.D.** and **Dr.L.Lakshmanan M.E., Ph.D. ,** Heads of the Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and a deep sense of gratitude to my Project Guide **Dr. T. Prem Jacob M.E., Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved the way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TABLE OF CONTENTS

ABSTRACT

In today's Social-media dominated world microblogging, Social networking and similar content-sharing platforms provides a platform(s) for global, social participation. Visibility of priorities at a national and international level for issues concerning disaster relief, climate change, political unrest and public health has significantly increased. These social media sites are almost always the fastest source for important news and information. Social media is a very popular medium for people to report on, especially in times of disasters and other incidents. In this paper, we go in-depth about a disaster identification system which utilizes social media as the primary source of information and utilizes the power of machine learning and Crowdsourcing.

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	EXPANSION
N.E.R	Named Entity Recognition
CNN	Convolutional Neural Network

CHAPTER 1

INTRODUCTION

The problem with traditional-media based news sources is that there is usually a time delay between the incident/disaster happening and news media outlets broadcasting said occurrence or writing articles about it. Social media sites like Twitter on the other hand, has its own set of problems which is that, the common user won't be able to see those vital pieces of information until it's "trending" or if it is "popular" enough.

Disaster-related posts on social media especially on Twitter contain lots of vital information like people injured, dead, missing/found, infrastructure and utility damage that can help rescue operations and disaster organizations to prioritize and improve the quality of their efforts. Machine learning is used to predict the future using historical data. It mainly focuses on the development of programs and models which have the ability to change when exposed to new and unknown data and it enables it to learn without being explicitly programmed. In this project, as we'll be handling a large volume of tweets, it is key to construct a Machine learning model that can detect these relevant tweets and extract data from them to improve the quality and response of rescue and relief efforts.

1.1 OVERVIEW

Microblogging can be seen as a form of lightweight chat which allows users to share short messages to the internet community. There are many popular microblogging services available, which includes Twitter, Plurk, Jaiku etc. Our focus for the purpose of this research paper is on Twitter, which allows its user to share short messages called Tweets which are essentially of 160 characters or less. These messages (tweets) can be sent and retrieved through a variety of means and front-end clients, including text messaging, e-mail, the web, and other third-party applications, which are enabled through Twitter's public API.

Microblogs are mainly used to share information and track general public opinion during any human-affecting event such as sports matches, political elections, natural disasters etc. Particularly in recent years, during the event of a natural disaster which usually occurs without any warning, social media has gained a lot of attention as an additional medium for crisis communication. Largely Twitter, being the most popular of microblogging websites, is being used by users to share news, photos and observations from the crisis site globally. Twitter being the real time data generated by the user community itself, is largely uncensored, easily accessible and is user-centric. Thus it provides an easier, comprehensive and economic analysis approach for users on fingertips. The hard problem is the sentiment analysis, specifically the behaviour and emotions that users express from natural disaster crisis sites to the distributed users across the globe.

The major motivation behind the proposed work is the fact that gathering information about people's emotions during and aftermath of a disaster is a task which is not yet automated and is mainly dependent upon manual surveys and interviews.

We propose to automate this task by making use of the data present on the social media websites. In this paper, we demonstrate an innovative application to analyze the effects of disaster on people and society through the use of twitter posts generated by users. The tool has ability to store the relevant (filtered) data from twitter and provides a detailed graphical analysis which encompasses user's emotions, disaster frequency and geographical distribution of several disasters such as earthquake, forest fire, floods, and droughts.

The main contributions of this work are:

•demonstration of geographical distribution of selected natural disaster within a given time period solely through the use of tweets.

• continent-wise occurrence frequency of selected natural disasters.

analysis of people's sentiment during a disaster by

applying sentiment-analysis on the tweet content

1.2 MACHINE LEARNING

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches.

In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

1.3 Machine Learning Strategies

In machine learning, tasks square measure typically classified into broad classes. These classes square measure supported however learning is received or however feedback on the educational is given to the system developed. Two of the foremost wide adopted machine learning strategies square measure supervised learning that trains algorithms supported example input and output information that's tagged by humans, and unattended learning that provides the algorithmic program with no tagged information so as to permit it to search out structure at intervals its computer file.

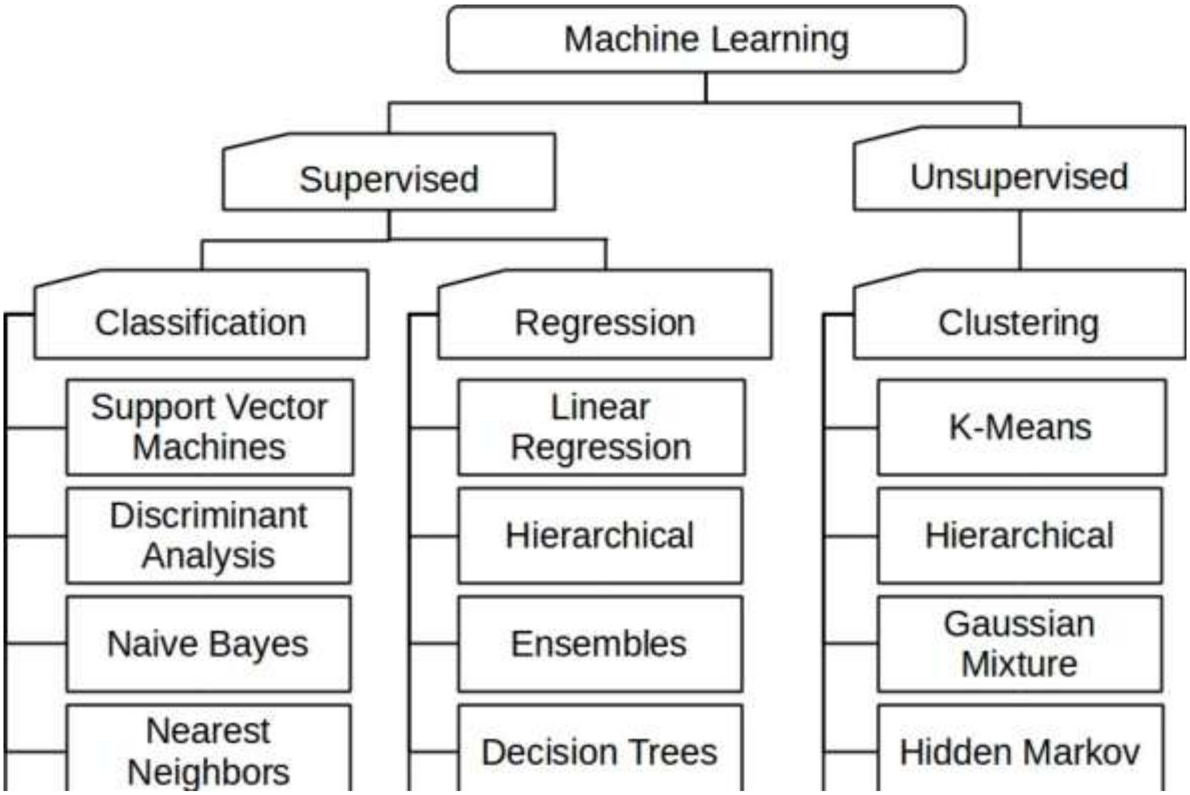


Fig 1.1 MACHINE LEARNING CLASSIFICATION

1.3.1 *Supervised Learning*

In supervised learning, the pc is given example inputs that square measure labelled with their desired outputs.The aim of this technique is for the algorithmic program to be ready to "learn" by comparison its actual output with the "taught" outputs to search out errors, and modify the model consequently. Supervised learning thus uses patterns to predict label values on extra unlabelled information. For example, with supervised learning, an algorithm may be fed data with images of sharks labelled as fish and images of oceans labelled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabelled shark images as fish and unlabelled ocean images as water.

A common use case of supervised learning is to use historical information to predict statistically probably future events. It's going to use historical stock exchange info to anticipate approaching fluctuations, or be used to filter spam emails. In supervised learning, labeled photos of dogs are often used as input file to classify unlabeled photos of dogs.

1.3.2 ***Unattended Learning***

In unattended learning, information is unlabeled, that the learning rule is left to seek out commonalities among its input file. The goal of unattended learning is also as easy as discovering hidden patterns at intervals a dataset, however it should even have a goal of feature learning, that permits the procedure machine to mechanically discover the representations that square measure required to classify data.

1.3.3 ***Unsupervised Learning***

Unsupervised learning is usually used for transactional information. You will have an oversized dataset of consumers and their purchases, however as a person's you'll probably not be able to add up of what similar attributes will be drawn from client profiles and their styles of purchases.

With this information fed into Associate in Nursing unattended learning rule, it should be determined that ladies of a definite age vary UN agency obtain unscented soaps square measure probably to be pregnant, and so a promoting campaign associated with physiological condition and baby will be merchandised.

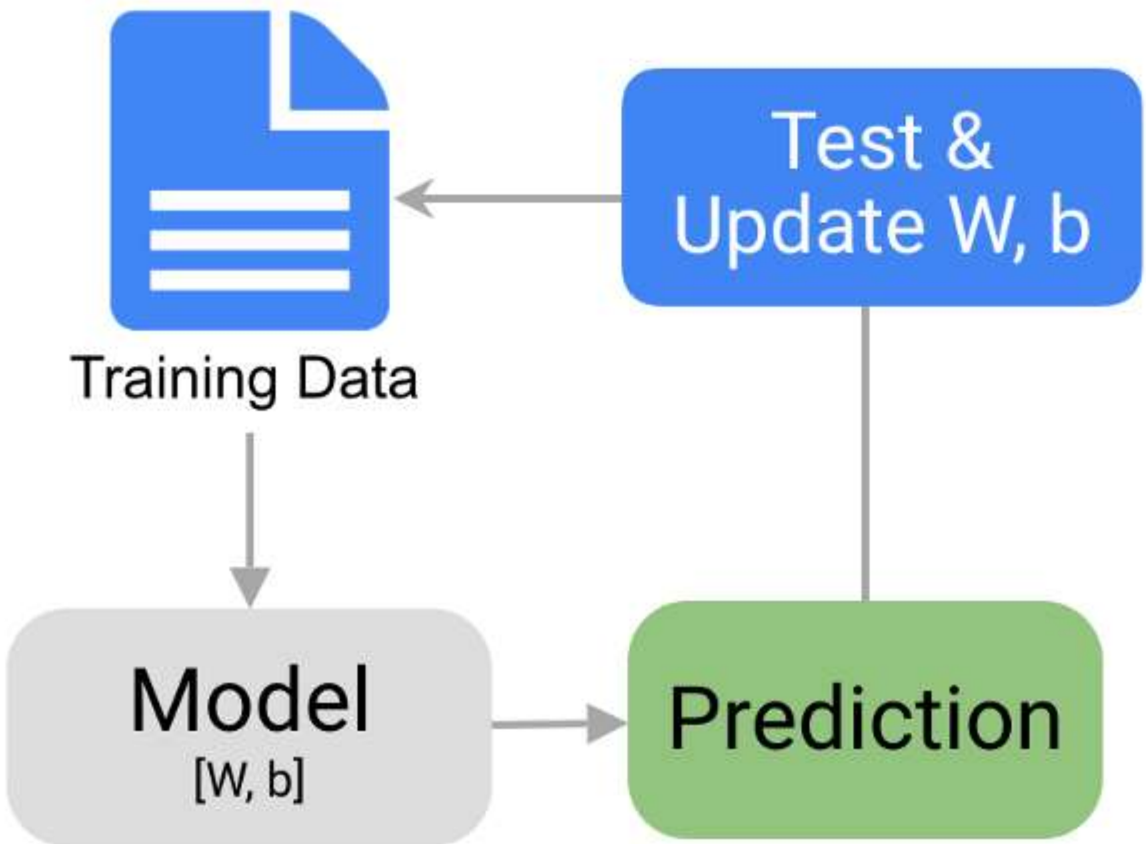


Fig 1.2 MACHINE LEARNING TASK

CHAPTER 2

LITERATURE SURVEY

2.1 RELATED WORK

[1] Yuya Shibuya (2017) - Mining Social Media for disaster management: leveraging social media data for community recovery.

The above paper talks about how managing and getting the information related to any disaster during a disaster is a serious concern.

[2] Jibo Xie & Tengfei Yang (2018) - Using Social Media Data to Enhance Disaster Response and Community service.

This paper sheds some light as to how disaster-related information is collected, managed, extracted, and classified from social media mediums for quicker disaster-related service.

[3] Shosuke Sato (2018) - Effectiveness and Limitations of Social Networking Services in Disaster responses.

Here the author Shosuke Sato discusses the uses of Social Networking Services in disaster responses. The effectiveness and restrictions of Social NEtworking Services are also talked about in this paper.

[4] Bernadette Joy Detera, Akira Kodaka & Kaya Onda (2021) - Twitter-based analysis of Disaster Sentiment during Typhoons and Earthquakes.

This paper talks about studying the various sentiments of people around the world during the history of numerous Typhoons and Earthquakes which can, in turn, improve the accuracy of identifying other types of disasters too.

[5] Malika Makker, Ramya B Ramanathan (2019) - Post Disaster Management using Satellite Imagery and social media.

In this paper, the authors talk about how important the use of Satellite Imagery is in both detecting and surveying areas, and especially those areas which have been affected thanks to the disasters.

[6] Huiji Gao & Rebecca Goolsby (2019) - Harnessing the Crowdsourcing power of social media for disaster relief.

This paper talks about the benefits and drawbacks of crowdsourcing applications that are implemented in disaster relief coordination.

[7] A.Mukkamala (2019) - Harnessing the Crowdsourcing Power of Social Media for Disaster Relief.

In this paper, the authors help us realize the scope and the obstacles that come along the way in using Social Media for Disaster relief management.

[8] Alex Lambert (2018) - Perspectives on Social Media and Communities in response and recovery. This paper talks about the collaborative study of disaster administration, the processing of information on social media w.r.t to disaster response and recuperation.

[9] S.Geetha & Vishnu Kumar Kaliappan (2018) Tweet Analysis Based on Distinct Opinion of Social Media Users.

The authors here talk about the potential for anticipating different tweet formats from twitter and improving the accuracy of detecting disaster-related tweets.

[10] Saurin R. Khedia, Shivam B. Parikh & Pradeep K. Atrey (2019) - A Framework to Detect Fake Tweet Images on Social Media.

The authors in this paper discuss using a framework that can help in detecting tweets that are either fake or have been meddled with, on any social media platform.

[11] Si Si Mar win & Than Nwe Aung (2017) - Target-oriented tweets monitoring system during natural disasters.

This paper institutes a tweet monitoring system that can help in recognizing messages/tweets that people can update during natural disasters into a group of disaster information-related categories and provide user desired target information spontaneously.

[12] Rabia Batool, Sungyoung Lee, Jahanzeb Maqbool (2013) - Precise tweet classification and sentiment analysis.

The authors of this paper examine tweets to categorize data and sentiments from Twitter even more accurately.

[13] Rasika Wagh, Payal Punde (2018) - Survey on Sentiment Analysis using Twitter Dataset.

This paper shows the various ways of analyzing sentiments from tweets, and also the various approaches went about to execute extraction of sentiments from tweets.

[14] Akash Kumar Gautam, Luv Mishra, Kush Mishra, Ajit Kumar, Shashwat Agarwal (2019) - Multimodal Analysis of Disaster Tweets. In this paper, the authors analyze several modes of data which are connected to natural disasters and classify them based on informative & non-informative.

[15] S. Devi, K Naveenkumar, S Shakti Ganesh (2021) - Location Based Twitter Emotion Classification for Disaster Management. This paper talks about designing a simple system that can identify and analyze tweets which can help towards disaster identification & recuperation.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

From the above literature survey, We have inferred that all the systems existing detect/identify disasters primarily through keywords, hashtags, and/or using only machine learning.

After implementing various algorithms, the tweets are identified and classified based on their disasters and is displayed on a Web-based app.

3.2 PROPOSED SYSTEM

We are proposing a system, which uses Machine Learning and Crowdsourcing. It is a multi-platform application which is efficient in notifying users. Our aim is to provide the masses with a disaster system which has high accuracy and which will also warn people about new disasters.

The primary domain we use in machine learning, in thate we will be using multiple pipelines which will process social media posts. To improve response time, we live-stream social media posts.

In this way, our system will be be less time consuming, give accurate predictions and will provide the masses with the latest information.

3.3 OBJECTIVE OF PROJECT

Developing a project based on machine learning (ML) and Crowdsourcing algorithms for detection of disasters can help in a more accurate detection than the conventional method is the main objective of the project. We have designed a Disaster Detection System using multiple piplines ad Machine Learning Models. Based on social media posts, the ML models give the output, i.e, the tweets and locations where the disaster has occured. This project helps people to be aware of disasters around them, while crowdsourcing enables people to give updates on situations.

3.4 SOFTWARE AND HARDWARE REQUIREMENTS

Web - app :

3.4.1 Technologies Used:

- Python 3
- Spacy
- Flask
- HTML
- CSS
- Javascript

- Java
- Docker (microservices)
- AWS EC2
- MongoDB

Mobile App :

- Kotlin
- Mongodb Realm

3.4.2 Software Requirements:

- Linux/Windows/Macos Operating System
- Docker, Docker-Compose

3.4.3 Hardware Requirements:

- Hard Disk : 80GB and above
- Ram : 4 GB and Above
- Processor : Pentium V and above

3.4.4 Libraries:

- ❖ **spacy** - Spacy is a library used for advanced Natural Language Processing in Python and Cython. It features state-of-the-art speed and neural network models for tagging, parsing, named entity recognition, text classification and more, multi-task learning with pre trained transformers like BERT, as well as a production ready training system and easy model packaging, deployment and workflow management.

- **PyMongo** - PyMongo is a Python distribution containing tools for working with MongoDB, and is the recommended way to work with MongoDB from Python.
- **Flask -** Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.
- **Flask_apscheduler -** APScheduler is a Flask extension which adds support for the APScheduler. where APScheduler is a Python library that lets you schedule your Python code to be executed later, either just once or periodically. You can add new jobs or remove old ones on the fly as you please.

3.5. PROGRAMMING LANGUAGES

3.5.1 *PYTHON*

Python is that the best programing language fitted to Machine Learning. In step with studies and surveys, Python is that the fifth most significant language yet because the preferred language for machine learning and information science. It's owing to the subsequent strengths that Python has –

- **Easy to be told and perceived-** The syntax of Python is simpler; thence it's comparatively straightforward, even for beginners conjointly, to be told and perceive the language.
- **Multi-purpose language −** Python could be a multi-purpose programing language as a result of it supports structured programming, object-oriented programming yet practical programming.
- **Support of the open supply community −** As being an open supply programming language, Python is supported by an awfully giant developer community. Because of this, the bugs square measure is simply mounted by the Python community. This characteristic makes Python terribly strong and adaptative.

3.5.2 DOMAIN

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches. In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. Thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

3.6. SYSTEM ARCHITECTURE

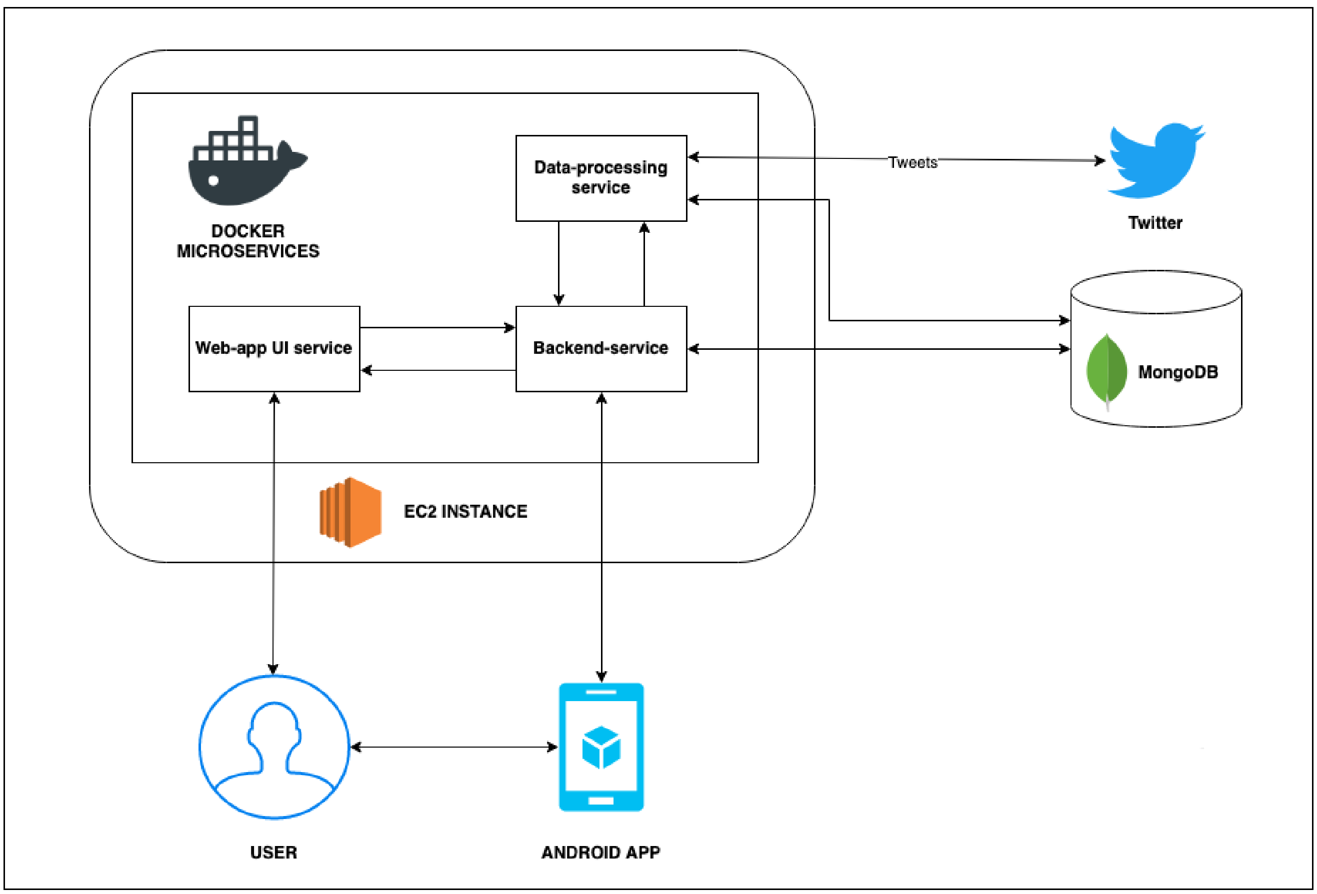


Fig 3.1 System Architecture

3.7 ALGORITHMS USED

3.7.1 CNN

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlap to cover the entire visual area.

3.7.2 NAIVE BAYES

It is a machine learning algorithm for classification problems which is based on Bayes theorem. The primary use of this is to do text classification. The Bayes theorem can be defined as:

$P(C|X)= P(X|C).P(C)\ P\ (X)$

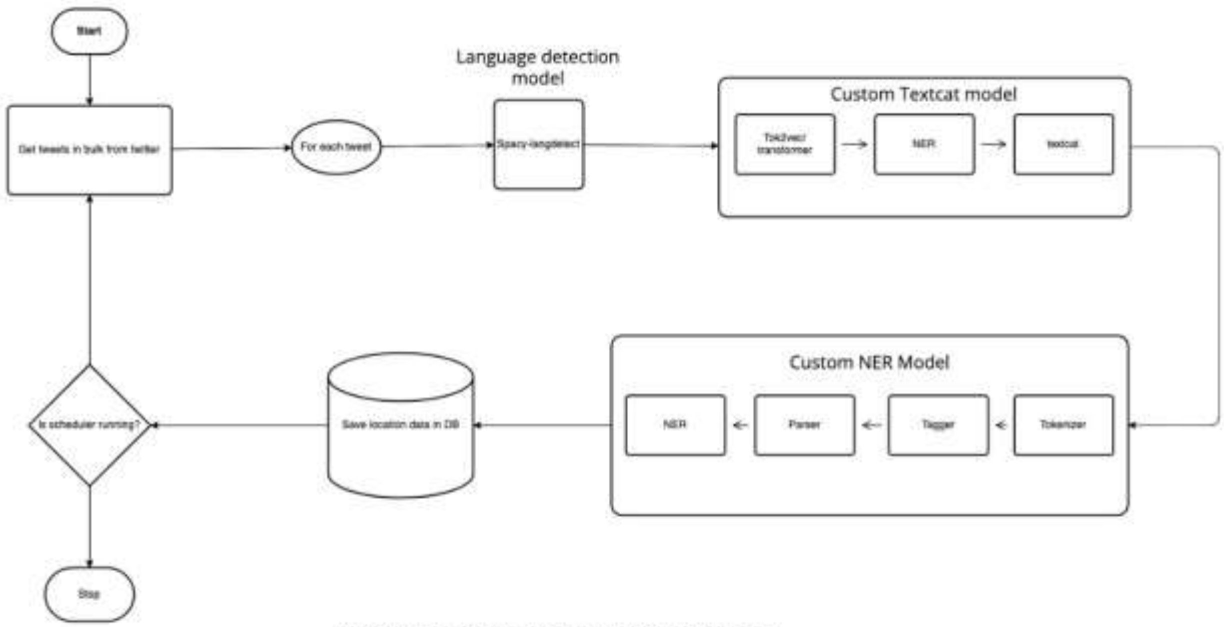
$P(C|X)$ is the probability of hypothesis C for the given data X. This is called the posterior probability.

$P(X|C)$ is the probability of data X given that the hypothesis C was true.

$P(C)$ is the probability of hypothesis C being true. This is called the prior probability of C.

11

P(X) is the probability of the data and evidence of data and is called marginal probability.



Workflow diagram for disaster detection system

Fig.3.2 Workflow Diagram

3.8.MODULES

3.8.1 DATA EXTRACTION :

This module deals with the extraction of tweets from the

web and storing them locally for analysis.

- Tweet Capturing: Data extraction was mainly done with the help of snscrape library. This library is capable of capturing tweets from the web and parsing them as String objects. These String objects can be further worked upon by the program for storage and manipulation.

3.8.2 Data Sorting :

Language Filtering : The captured tweets goes through a language detection pipeline (lang-detect) which filters out non-English tweets which are then discarded. Here only tweets that are in English are allowed to move on.

- Text-classification :The captured tweets are now analyzed and their disaster category is determined. This is done by checking each tweet against a set of predefined weighted keywords.. The tweets which do not belong to any category are readily discarded.

3.8.3 Analysis of Data :

The categorized data is then passed on to the third module for analysis and interpretation.

- Geo-Tagging: The location of disaster is the foremost matter of interest in any type of disaster analysis. But in many cases tweets are posted by people who are not actually tweeting from the affected region. We extract the location data from the tweets by using a custom NER Model on the tweets to find the location mentioned. We then use Mapbox API to find the co-ordinates of the extracted locations..

3.8.4 Crowdsourcing

All the collected data is then visualized and mapped in a interactive user interface on a web-app and a mobile app. In these applications, the user also has the ability to add locations of disasters and also add other important information to new/existing disasters.

CHAPTER 4

RESULTS AND DISCUSSION

Our project mainly focuses on designing a Disaster Detection System that uses social media as its primary data source and a system that also harnesses the power of machine learning and crowdsourcing. This project provides a multi-platform app where users can view the data visualized as a map. The user can also verify, update and add vital information related to a disaster. This crowdsourcing element increases the accuracy and provides vital information especially from a few eyes on the ground (No. of people in need of help, pictures, videos). This system also can alert the user about any disasters/incidents happening in their area. The multiple Machine learning models and pipelines we used and the entire system as a whole is a one-stop solution for the masses to view real-time information about disasters and get alerted about any disaster-related events happening nearby.

Future works may include addition of features like allowing users to upload photos and videos of affected sites, and advanced filtering of crowdsourced information, image classification of social media posts and inclusion of multiple social media platforms as a data source.

CHAPTER 5

CONCLUSION

We have created a disaster identification system that uses Twitter as the primary source of information. This system uses a language detection pipeline to filter out non-English tweets, uses a custom Text classification model to classify disaster-related tweets, and finally, this system uses a custom NER (Named Entity Recognition Model) to extract vital information (geo-location data) from them. All the extracted data can be verified and appended by the masses with the help of a crowdsourcing component. Future work would include expanding the data sources to include other social media platforms, news-sites and extracting multi-media content related to disasters from them.

REFERENCES

[1] V. Nunavath and M. Goodwin, "The Role of Artificial Intelligence in Social Media Big data Analytics for Disaster Management -Initial Results of a Systematic Literature Review," 2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2018, pp. 1-4, doi: 10.1109/ICT-DM.2018.8636388.

[2] J. Xie and T. Yang, "Using Social Media Data to Enhance Disaster Response and Community Service," 2018 International Workshop on Big Geospatial Data and Data Science (BGDDS), 2018, pp. 1-4, doi: 10.1109/BGDDS.2018.8626839.

[3] S. Sato, "Effectiveness and Limitations of Social Networking Services in Disaster Responses: A Review 7 Years on from the 2011 Great East Japan Earthquake," 2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2018, pp. 1-6, doi: 10.1109/ICT-DM.2018.8636369.

[4] L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan and L. Shuib, "Automated Machine Learning Approaches for Emergency Response and Coordination via Social Media in the Aftermath of a Disaster: A Review," in IEEE Access, vol. 9, pp. 68917-68931, 2021, doi: 10.1109/ACCESS.2021.3074819.

[5] M. R. Sumalatha, P. Batsa, A. Sinha and P. Shrinath, "Social media for disaster relief — Geo distributed social service system," 2015 Seventh International Conference on Advanced Computing (ICoAC), 2015, pp. 1-6, doi: 10.1109/ICoAC.2015.7562783.

[6] M. Makker, R. Ramanathan and S. B. Dinesh, "Post Disaster Management Using Satellite Imagery and Social Media Data," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2019, pp. 1-6, doi: 10.1109/CSITSS47250.2019.9031042.

[7] H. Gao, G. Barbier and R. Goolsby, "Harnessing the Crowdsourcing Power of Social Media for Disaster Relief," in IEEE Intelligent Systems, vol. 26, no. 3, pp. 10-14, May-June 2011, doi: 10.1109/MIS.2011.52.

[8] Y. Shibuya, "Mining social media for disaster management: Leveraging social media data for community recovery," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3111-3118, doi: 10.1109/BigData.2017.8258286.

[9] S. Z. Razavi and M. Rahbari, "Understanding Reactions to Natural Disasters: a Text Mining Approach to Analyze Social Media Content," 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), 2020, pp. 1-7, doi: 10.1109/SNAMS52053.2020.9336570.

[10] K. Aziz, D. Zaidouni and M. Bellafkih, "Social Network Analytics: Natural Disaster Analysis Through Twitter," 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), 2019, pp. 1-7, doi: 10.1109/ICDS47004.2019.8942337.

[11] C. Fan, F. Wu and A. Mostafavi, "A Hybrid Machine Learning Pipeline for Automated Mapping of Events and Locations From Social Media in Disasters," in IEEE Access, vol. 8, pp. 10478-10490, 2020, doi: 10.1109/ACCESS.2020.2965550.

[12] A. Mukkamala and R. Beck, "Social media for disaster situations: Methods, opportunities and challenges," 2017 IEEE Global Humanitarian Technology Conference (GHTC), 2017, pp. 1-9, doi: 10.1109/GHTC.2017.8239277.

[13] B. Anbalagan and C. Valliyammai, "#ChennaiFloods: Leveraging Human and Machine Learning for Crisis Mapping during Disasters Using Social Media," 2016 IEEE 23rd International Conference on High Performance Computing Workshops (HiPCW), 2016, pp. 50-59, doi: 10.1109/HiPCW.2016.016.

[14] B. Abedin, A. Babar and A. Abbasi, "Characterization of the Use of Social Media in Natural Disasters: A Systematic Review," 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, 2014, pp. 449-454, doi: 10.1109/BDCloud.2014.17.

APPENDICES

A) SOURCE CODE

Text classification and GeoLocation data extraction:

```
def get_latest_tweets():
    global nlp
    date = datetime.now()

    date = date.strftime("%Y-%m-%d")

      os.system(f"snscrape --jsonl --max-results 1000 --since {date} twitter-search 'killed OR dead OR
disaster OR tragedy OR incident AND india' > text-query-tweets.json")



    tweets = []
    for line in open('text-query-tweets.json', 'r'):
        tweet = json.loads(line)

        text_content = tweet["content"]
        doc = nlp(text_content)
        detect_language = doc._.language
        english_flag = False



        if detect_language["language"] == 'en' and detect_language['score'] >= 0.85:
            tweets.append(tweet["content"])

    nlp2 = en_textcat_demo.load()

    docs = list(nlp2.pipe(tweets))
```

```
    nlp_updated = spacy.load(output_dir)

    result = []
    locations_list = []

    for doc in docs:

        if doc.cats["DISASTER"] > 0.75:

            print(doc.text)
            loc_doc = nlp_updated(doc.text)
            for ent in loc_doc.ents:
                if ent.label_ == "LOC":
                    print(ent.label)
                    locations_list.append(str(ent))
```

Get coords and write to DB :

```
def write_locations_db(locations_list):

    collection = tweets_db["location_hits"]

    current_datetime = datetime.now()

    for location in locations_list:
```

```
    URL =
f"https://api.mapbox.com/geocoding/v5/mapbox.places/{location}.json?access_token=pk.ey"


    r = requests.get(url = URL)
    data = r.json()
    try:
      coords = data["features"][0]['center']
    except:
      coords = []
    check_location = collection.find_one({"location": location})


    if check_location :
      hits = check_location["count"]
      collection.update_one({"location": location},{"$set": {"count": hits + 1, "last_hit" :
current_datetime}}) # need to change current datetime to latest tweet
    else :


      location_hit_entry = {
        "location" : location,
        "count" : 1,
        "coords" : coords,
        "disaster_type"   : None ,
        "additional-info" : None,
        "last_hit" : current_datetime # need to change current datetime to latest tweet,
      }
      collection.insert_one(location_hit_entry)
  return locations_list
```

Training NER Model

```
import spacy, json
nlp=spacy.load("en_core_web_sm")

ner=nlp.get_pipe('ner')



LABEL = "LOCATION"

TRAIN_DATA =[ ]
with open('./location_training.jsonl', 'r') as json_file:
    json_list = list(json_file)

for json_str in json_list:
    result = json.loads(json_str)
    training_data_entry = result["data"]
    training_data_entry[1]["entities"][0] = tuple(training_data_entry[1]["entities"][0])
    training_data_entry = tuple(training_data_entry)
    TRAIN_DATA.append(training_data_entry)
print(TRAIN_DATA)
ner.add_label(LABEL)

optimizer = nlp.resume_training()
move_names = list(ner.move_names)

pipe_exceptions = ["ner", "trf_wordpiecer", "trf_tok2vec"]
```

```
other_pipes = [pipe for pipe in nlp.pipe_names if pipe not in pipe_exceptions]



from spacy.util import minibatch, compounding
import random

with nlp.disable_pipes(*other_pipes) :

  sizes = compounding(1.0, 4.0, 1.001)

  for itn in range(30):
    random.shuffle(TRAIN_DATA)

    batches = minibatch(TRAIN_DATA, size=sizes)    losses = {}
    for batch in batches:
      texts, annotations = zip(*batch)

      nlp.update(texts, annotations, sgd=optimizer, drop=0.35, losses=losses)
      print("Losses", losses)



from pathlib import Path
output_dir=Path('./location_model')



if not output_dir.exists():
```

```
  output_dir.mkdir()
nlp.meta['name'] = 'my_ner'  nlp.to_disk(output_dir)
print("Saved model to", output_dir)



print("Loading from", output_dir)
nlp2 = spacy.load(output_dir)
assert nlp2.get_pipe("ner").move_names == move_names
doc2 = nlp2(' Dosa is an extremely famous south Indian dish')
for ent in doc2.ents:
  print(ent.label_, ent.text)

print("Loading from", output_dir)
nlp_updated = spacy.load(output_dir)
doc = nlp_updated("Fridge can be ordered in FlipKart" )
print("Entities", [(ent.text, ent.label_) for ent in doc.ents])
```

B) SCREENSHOTS

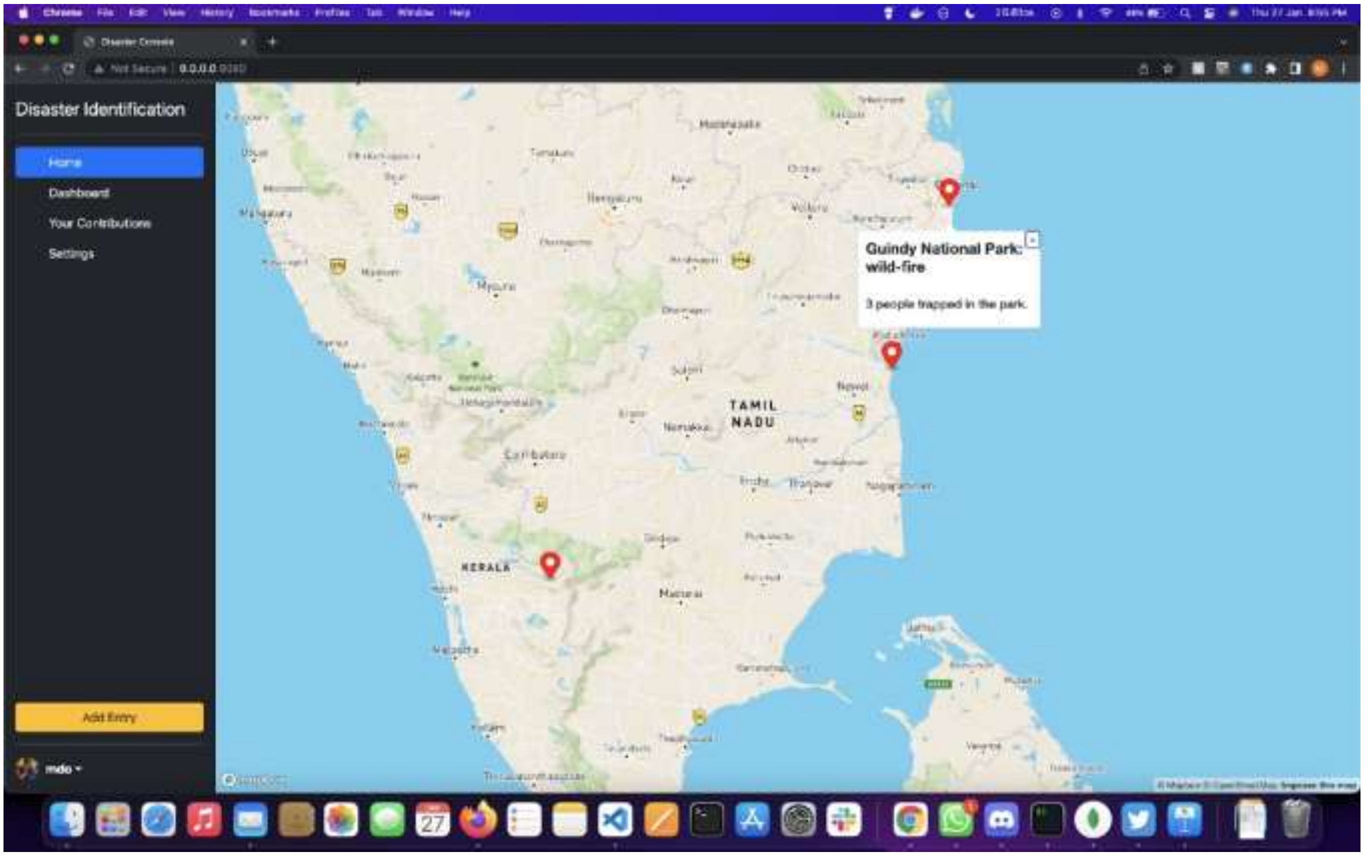


Fig 4.1 LOCATION OF DISASTERS ON WEB-APP

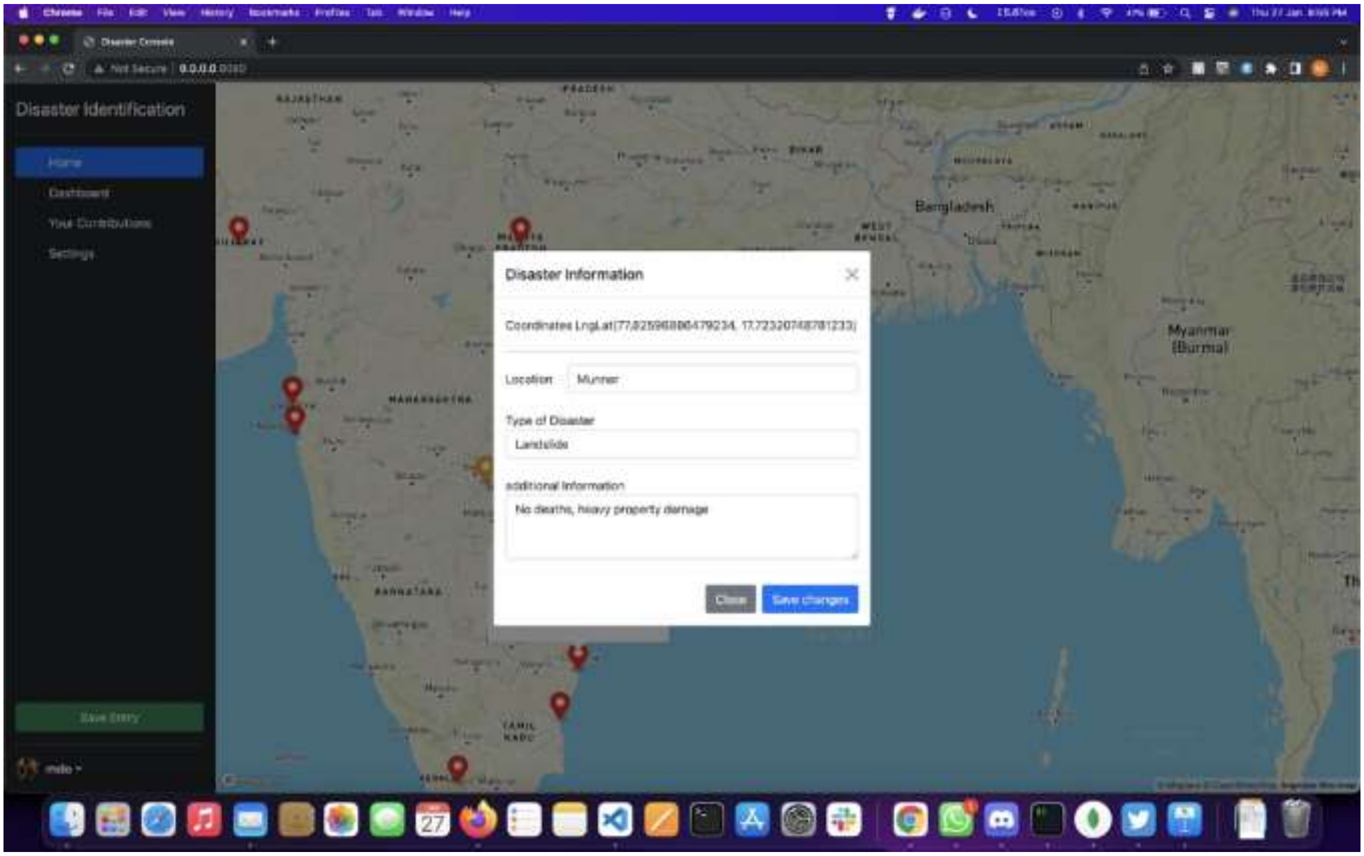


Fig 4.2 CROWDSOURCING ELEMENT

Fig 4.3 Disaster Locations in the mobile app

C) PLAGIARISM REPORT

PaperPass.net

Title

DISASTER DETECTION SYSTEM USING SOCIAL MEDIA, MACHINE LEARNING, AND CROWDSOURCING

5%	5%	0%
SIMILARITY INDEX	ACADEMIC	INTERNET

Date:	2022-02-28 09:53:04(+00:00 UTC)
Report ID:	621c9b8c062799840
Word count:	2345
Character count:	13161

Similar sources		
1	• Spatial and temporal analysis of Twitter • David Goergen,Angelo Migliosi,Vijay K. Gurbani,Radu State,Thomas Engel • Proceedings of the Conference on Principles, Systems and Applications of IP Telecommunications - IPTComm '14,2014 Others	0.4%
2	• Melanoma and sarcomas • European Journal of Surgical Oncology (EJSO),2015 Academic	0.3%
3	• General oncology • European Journal of Surgical Oncology (EJSO),2015 Academic	0.3%
4	• SenseTag: A tagging tool for constructing Thai sentiment lexicon • Kanokorn Trakultaweekoon,Supon Klaithin • 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE),2016 Others	0.3%
5	• Development of the Student Refuge Information System Using IC Card Type Student Identification Cards • Hiroaki Yuze,Yuji Aota,Takashi Kawada • 2009 International Conference on Advanced Information Networking and Applications Work-shops,2009 Others	0.3%
6	• IAU volume 7 issue S278 Cover and Front matter • Proceedings of the International Astronomical Union,2011 Academic	0.3%
7	• Back Matter for Volume 1347 Others	0.3%
8	• Cancer of the esophagus and stomach • European Journal of Surgical Oncology (EJSO),2015 Academic	0.3%
9	• From online via offline to online: how online visibility of tourism information shapes and is shaped by offline visits • Mi Kyung Lee,Ho Young Yoon,Han Woo Park • Journal of Travel & Tourism Marketing,2017 Academic	0.3%
10	• ECF report • 2019 International Conference on Wireless Networks and Mobile Communications (WINCOM),2019 Others	0.2%
11	• Crime rate detection using social media of different crime locations and Twitter part-of-speech tag-ger with Brown clustering • Thanh Vo,Rohit Sharma,Raghvendra Kumar,Le Hoang Son,Binh Thai Pham,Dieu Tien Bui,Ishaani P... • Journal of Intelligent & Fuzzy Systems,2020 Academic	0.2%
12	• A Generic Architecture for a Social Network Monitoring and Analysis System • Alexander Semenov,Jari Veijalainen,Alexander Boukhanovsky • 2011 14th International Conference on Network-Based Information Systems,2011 Others	0.2%

13	• Preface • John A Parrotta • International Forestry Review,2014 Academic	0.2%
14	• Theme A: Forests for People • International Forestry Review,2014 Academic	0.2%
15	• Examining the impacts of mental workload and task-technology fit on user acceptance of the social media search system • Yan (Mandy) Dang,Yulei (Gavin) Zhang,Susan A Brown,Hsinchun Chen • Information Systems Frontiers,2018 Academic	0.2%
16	• Cancer of the liver, pancreas and biliary ducts • European Journal of Surgical Oncology (EJSO),2015 Academic	0.2%
17	• The Gravity Probe B test of general relativity • C W F Everitt,B Muhlfelder,D B DeBra,B W Parkinson,J P Turneaure,A S Silbergleit,E B Acwor... • Classical and Quantum Gravity,2015 Academic	0.2%
18	• Gynecologic oncology • European Journal of Surgical Oncology (EJSO),2015 Academic	0.2%
19	• A humanistic approach to the study of social media • Ashley Rose Kelly,Meagan Kittle Autry • Proceedings of the 29th ACM international conference on Design of communication - SIGDOC '11,2011 Others	0.2%
20	• Riskr: a web 2.0 platform to monitor and share disaster information • Jess Farber,Trina Myers,Jarrod Trevathan,Ian Atkinson,Trevor Andersen • International Journal of Grid and Utility Computing,2015 Academic	0.2%
21	• Social Media Competitive Analysis of Shoe Brands on Customer Experiences • Imran Ahmad,Tahir Ali,Asad Nazir,Shahid Kamal • Communications in Computer and Information Science Intelligent Technologies and Applications,2019 Others	0.2%
22	• Taming Big Data • Christopher A Bail • Sociological Methods & Research,2016 Academic	0.2%
23	• Government 2.0 • Alexander R.M. Schellong • Transforming Government: People, Process and Policy,2008 Academic	0.2%
24	• Content mining framework in social media: A FIFA world cup 2014 case analysis • Guilherme M. Thomaz,Alexandre A. Biz,Eduardo M. Bettoni,Luiz Mendes-Filho,Dimitrios Buhali... • Information & Management,2017 Academic	0.2%
25	• Analyzing the role of media orchestration in conducting disinformation campaigns on blogs • Kiran Kumar Bandeli,Nitin Agarwal • Computational and Mathematical Organization Theory,2018	0.2%

Academic

26	• Don't You Dare Push Me: How Persuasive Social Media Tactics Shape Customer Engagement • Welf H. Weiger,Maik Hammerschmidt,Hauke A. Wetzel • Journal of the Association for Consumer Research,2018 Academic	0.2%
27	• Mining photo-sharing websites to study ecological phenomena • Haipeng Zhang,Mohammed Korayem,David J. Crandall,Gretchen LeBuhn • Proceedings of the 21st international conference on World Wide Web - WWW '12,2012 Others	0.2%
28	• Active Management of Operational Risk in the Regimes of the "Unknown": What Can Machine Learning or Heuristics Deliver? • Udo Milkau,Jürgen Bott • Risks,2018 Academic	0.2%
29	• Social-media-based public policy informatics: Sentiment and network analyses of U.S. Immigration and border security • Wingyan Chung,Daniel Zeng • Journal of the Association for Information Science and Technology,2015 Academic	0.2%
30	• Sub-event detection during natural hazards using features of social media data • Dhekar Abhik,Durga Toshniwal • Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion,2013 Others	0.2%

D) JOURNAL PAPER

DISASTER DETECTION SYSTEM USING SOCIAL MEDIA, MACHINE LEARNING, AND CROWDSOURCING

Ajesh M 1, Akhil Anand 2, Dr.T.Prem Jacob3

1,2UG Student, Department of Computer Science and Engineering

Sathyabama Institute of Science and Technology

3Associate Professor, Department of Computer Science and Engineering

Sathyabama Institute of Science and Technology

1ajeshmartin01@gmail.com, 2akhiltuo9@gmail.com, 3premjac@yahoo.com

Abstract - In today's Social-media dominated world, microblogging, social networking, and similar content-sharing platforms provide a platform(s) for global, social participation. Visibility of priorities at a national and international level for issues concerning disaster relief, climate change, political unrest, and public health has significantly increased. These social media sites are almost always the fastest source for important news and information. Social media is a very popular medium for people to report on, especially in times of disasters and other incidents. In this paper, we go in-depth about a disaster identification system which utilizes social media as the primary source of information and utilizes the power of machine learning and Crowdsourcing."

Keywords: Spacy, Snscrape, N.E.R, Spacy-Langdetect, Crowdsourcing.

I. INTRODUCTION

The problem with traditional-media-based news sources is that there is usually a time delay between the incident/disaster happening and news media outlets broadcasting said occurrence or writing articles about it. Social media sites like Twitter, on the other hand, have their own set of problems which is that, the common user won't be able to see those vital pieces of information until it's "trending" or if it is "popular" enough.

Disaster-related posts on social media especially on Twitter contain lots of vital information like people injured, dead, missing/found, infrastructure, and utility damage that can help rescue operations and disaster organizations to prioritize and improve the quality of their efforts. Machine learning is used to predict the future using historical data. It mainly focuses on the development of programs and models which can change when exposed to new and unknown data and it enables it to learn without being explicitly programmed. In this project, as we'll be handling a large volume of tweets, it is key to construct a Machine learning model that can detect these relevant tweets and extract data from them to improve the quality and response of rescue and relief efforts.

II. PROPOSED WORK

We aim to solve the time delay of traditional news media and the volume of social media tweets/posts problems by live-streaming tweets and using machine learning to detect disasters by text analysis and classification.

We use machine learning models which aid in filtering out disaster-related tweets and extracting data from them. Our work also includes a crowdsourcing component where users can verify information about a disaster and add other useful information to it. The extracted and crowdsourced data, which may contain the location and other vital information of said disaster, can be viewed on a user-friendly web app or a mobile app. The multi-platform app allows users to be alerted to any disaster/incidents in their area. The Crowdsourcing component further decreases

III. WORKING OF PROJECT

A. Tweet(s) Retrieval using Snscrape :

The first step in our system is to collect data from social media sites. We use Twitter as the main source of information. We periodically retrieve large amounts of tweets that were tweeted within a certain period using Snscrape (a python package that can return posts/tweets from multiple social media platforms). We retrieve tweets using a large number of keywords to reduce the volume of unrelated tweets. We also don’t limit tweets by geographical boundaries.

B. Tweet Filtering and Classification :

We use Spacy-long detect (a fully customizable language detection pipeline) which filters out tweets that are in other languages (English being the primary language) to increase the text classification accuracy in the next step. Then we directly feed those tweets into a custom text classification model built with Spacy to classify disaster-related tweets.

C. Geo-location data extraction :

We then extract relevant data (location) from said tweets using a custom NER (Named Entity Recognition) model built with Spacy. We store all the relevant data in a database.

D. End-user Interaction and Crowdsourcing :

The end-user can view the disasters/incidents and their details on an interactive map on two platforms (Web and Mobile). Here we also integrate a cloud sourcing element where users can verify, update and add vital information, especially information from people who are in and around a disaster-affected area (Pictures, No.of people in need of help, etc). The end-user can also save their location(s) to receive alerts about disasters or incidents around them.

IV. ARCHITECTURE DIAGRAM

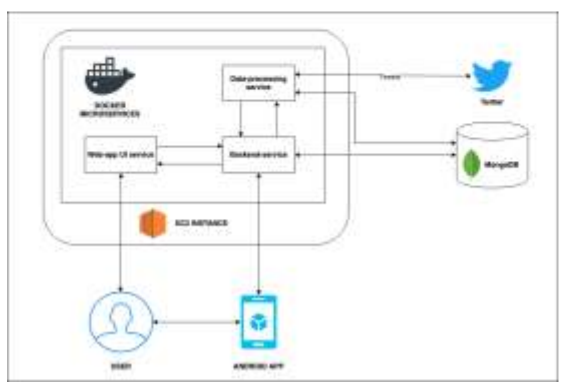


Fig.1 : Architecture of the proposed system

All the processing and the Web-app itself is packaged, deployed, and is run using docker containers on a microservices architecture.

It consists of three different services, the Data-processing service, which uses all the machine learning models to detect tweets on disasters. The backend service serves data retrieval requests for the web app (UI service) and the mobile application. And there's the web app itself, this service is the actual web app the end-user can interact with.

All the above services run on auto-scalable EC2 Instances, which means the admin doesn't have to manually provision additional servers in case of unpredictable high traffic (which in our use-case is highly probable). The data extracted is stored in a MongoDB cloud cluster.

V. LITERATURE REVIEW

[1] "Yuya Shibuya (2017) - Mining Social Media for disaster management: leveraging social media data for community recovery."

The above paper talks about how managing and getting the information related to any disaster during a disaster is a serious concern.

[2]"Jibo Xie & Tengfei Yang (2018) - Using Social Media Data to Enhance Disaster Response and Community service. "

This paper sheds some light as to how disaster-related information is collected, managed, extracted, and classified from social media mediums for quicker disaster-related service."

[3]"Shosuke Sato (2018) - Effectiveness and Limitations of Social Networking Services in Disaster responses.

Here the author Shosuke Sato discusses the uses of Social Networking Services in disaster responses. The effectiveness and restrictions of Social NEtworking Services are also talked about in this paper.""""

[4]"Bernadette Joy Detera, Akira Kodaka & Kaya Onda (2021) - Twitter-based analysis of Disaster Sentiment during Typhoons and Earthquakes.

This paper talks about studying the various sentiments of people around the world during the history of numerous Typhoons and Earthquakes which can, in turn, improve the accuracy of identifying other types of disasters too."

[5]"Malika Makker, Ramya B Ramanathan (2019) - Post Disaster Management using Satellite Imagery and social media.

In this paper, the authors talk about how important the use of Satellite Imagery is in both detecting and surveying areas, and especially those areas which have been affected thanks to the disasters. "

[6]"Huiji Gao & Rebecca Goolsby (2019) - Harnessing the Crowdsourcing power of social media for disaster relief.

This paper talks about the benefits and drawbacks of crowdsourcing applications that are implemented in disaster relief coordination. "

[7]"A.Mukkamala (2019) - Harnessing the Crowdsourcing Power of Social Media for Disaster Relief.

In this paper, the authors help us realize the scope and the obstacles that come along the way in using Social Media for Disaster relief management."

[8]"Alex Lambert (2018) - Perspectives on Social Media and Communities in response and recovery. This paper talks about the collaborative study of disaster administration, the processing of information on social media w.r.t to disaster response and recuperation."

[9]"S.Geetha & Vishnu Kumar Kaliappan (2018) Tweet Analysis Based on Distinct Opinion of Social Media Users.

The authors here talk about the potential for anticipating different tweet formats from twitter and improving the accuracy of detecting disaster-related tweets.”

[10]”Saurin R. Khedia, Shivam B. Parikh & Pradeep K. Atrey (2019) - A Framework to Detect Fake Tweet Images on Social Media.

The authors in this paper discuss using a framework that can help in detecting tweets that are either fake or have been meddled with, on any social media platform.”

[11]”Si Si Mar win & Than Nwe Aung (2017) - Target-oriented tweets monitoring system during natural disasters.

This paper institutes a tweet monitoring system that can help in recognizing messages/tweets that people can update during natural disasters into a group of disaster information-related categories and provide user desired target information spontaneously.”

[12]”Rabia Batool, Sungyoung Lee, Jahanzeb Maqbool (2013) - Precise tweet classification and sentiment analysis.

The authors of this paper examine tweets to categorize data and sentiments from Twitter even more accurately.”

[13]”Rasika Wagh, Payal Punde (2018) - Survey on Sentiment Analysis using Twitter Dataset.

This paper shows the various ways of analyzing sentiments from tweets, and also the various approaches went about to execute extraction of sentiments from tweets.”

[14]”Akash Kumar Gautam, Luv Mishra, Kush Mishra, Ajit Kumar, Shashwat Agarwal (2019) - Multimodal Analysis of Disaster Tweets. In this paper, the authors analyze several modes of data which are connected to natural disasters and classify them based on informative & non-informative.”

[15]”S. Devi, K Naveenkumar, S Shakti Ganesh (2021) - Location Based Twitter Emotion Classification for Disaster Management. This paper talks about designing a simple system that can identify and analyze tweets which can help towards disaster identification & recuperation.

VI. RESULT

Fig. 2 : trending disasters on social media (2021-2022)

Our project mainly focuses on designing a Disaster Detection System that uses social media as its primary data source and a system that also harnesses the power of machine learning and crowdsourcing. This project provides a multi-platform app where users can view the data visualized as a map. The user can also verify, update and add vital information related to a disaster. This crowdsourcing element

increases the accuracy and provides vital information.

Fig. 3 : Location of disasters on the Web UI

Fig 4 : The crowdsourcing feature

Information especially from eyes on the ground (No. of people in need of help, pictures, videos). This system also can alert the user about any disasters/incidents happening in their area. This system is a one-stop solution for the masses to view real-time information about disasters and get alerted about any disaster-related events happening nearby.

VII. CONCLUSION

We have created a disaster identification system that uses Twitter as the primary source of information. This system uses a language detection pipeline to filter out non-English tweets, uses a custom Text classification model to classify disaster-related tweets, and finally, this system uses a custom NER (Named Entity Recognition Model) to extract vital information (geo-location data) from them. All the extracted data can be verified and appended by the masses with the help of a crowdsourcing component. Future work would include expanding the data sources to include other social media platforms, news-sites and extracting multi-media content related to disasters from them.

VIII. REFERENCES

[1] V. Nunavath and M. Goodwin, "The Role of Artificial Intelligence in Social Media Big data Analytics for Disaster Management -Initial Results of a Systematic Literature Review," 2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2018, pp. 1-4, doi: 10.1109/ICT-DM.2018.8636388.

[2] J. Xie and T. Yang, "Using Social Media Data to Enhance Disaster Response and Community Service," 2018 International Workshop on Big Geospatial Data and Data Science (BGDDS), 2018, pp. 1-4, doi: 10.1109/BGDDS.2018.8626839.

[3] S. Sato, "Effectiveness and Limitations of Social Networking Services in Disaster Responses: A Review 7 Years on from the 2011 Great East Japan Earthquake," 2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2018, pp. 1-6, doi: 10.1109/ICT-DM.2018.8636369.

[4] L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan and L. Shuib, "Automated Machine Learning Approaches for Emergency Response and Coordination via Social Media in the Aftermath of a Disaster: A Review," in IEEE Access, vol. 9, pp. 68917-68931, 2021, doi: 10.1109/ACCESS.2021.3074819.

[5] M. R. Sumalatha, P. Batsa, A. Sinha and P. Shrinath, "Social media for disaster relief — Geo distributed social service system," 2015 Seventh International Conference on Advanced Computing (ICoAC), 2015, pp. 1-6, doi: 10.1109/ICoAC.2015.7562783.

[6] M. Makker, R. Ramanathan and S. B. Dinesh, "Post Disaster Management Using Satellite Imagery and Social Media Data," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2019, pp. 1-6, doi: 10.1109/CSITSS47250.2019.9031042.

[7] H. Gao, G. Barbier and R. Goolsby, "Harnessing the Crowdsourcing Power of Social Media for Disaster Relief," in IEEE Intelligent Systems, vol. 26, no. 3, pp. 10-14, May-June 2011, doi: 10.1109/MIS.2011.52.

[8] Y. Shibuya, "Mining social media for disaster management: Leveraging social media data for community recovery," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3111-3118, doi: 10.1109/BigData.2017.8258286.

[9] S. Z. Razavi and M. Rahbari, "Understanding Reactions to Natural Disasters: a Text Mining Approach to Analyze Social Media Content," 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), 2020, pp. 1-7, doi: 10.1109/SNAMS52053.2020.9336570.

[10] K. Aziz, D. Zaidouni and M. Bellafkih, "Social Network Analytics: Natural Disaster Analysis Through Twitter," 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), 2019, pp. 1-7, doi: 10.1109/ICDS47004.2019.8942337.

[11] C. Fan, F. Wu and A. Mostafavi, "A Hybrid Machine Learning Pipeline for Automated Mapping of Events and Locations From Social Media in Disasters," in IEEE Access, vol. 8, pp. 10478-10490, 2020, doi: 10.1109/ACCESS.2020.2965550.

[12] A. Mukkamala and R. Beck, "Social media for disaster situations: Methods, opportunities and challenges," 2017 IEEE Global Humanitarian Technology Conference (GHTC), 2017, pp. 1-9, doi: 10.1109/GHTC.2017.8239277.

[13] B. Anbalagan and C. Valliyammai, "#ChennaiFloods: Leveraging Human and Machine Learning for Crisis Mapping during Disasters Using Social Media," 2016 IEEE 23rd International Conference on High Performance Computing Workshops (HiPCW), 2016, pp. 50-59, doi: 10.1109/HiPCW.2016.016.

[14] B. Abedin, A. Babar and A. Abbasi, "Characterization of the Use of Social Media in Natural Disasters: A Systematic Review," 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, 2014, pp. 449-454, doi: 10.1109/BDCloud.2014.17.