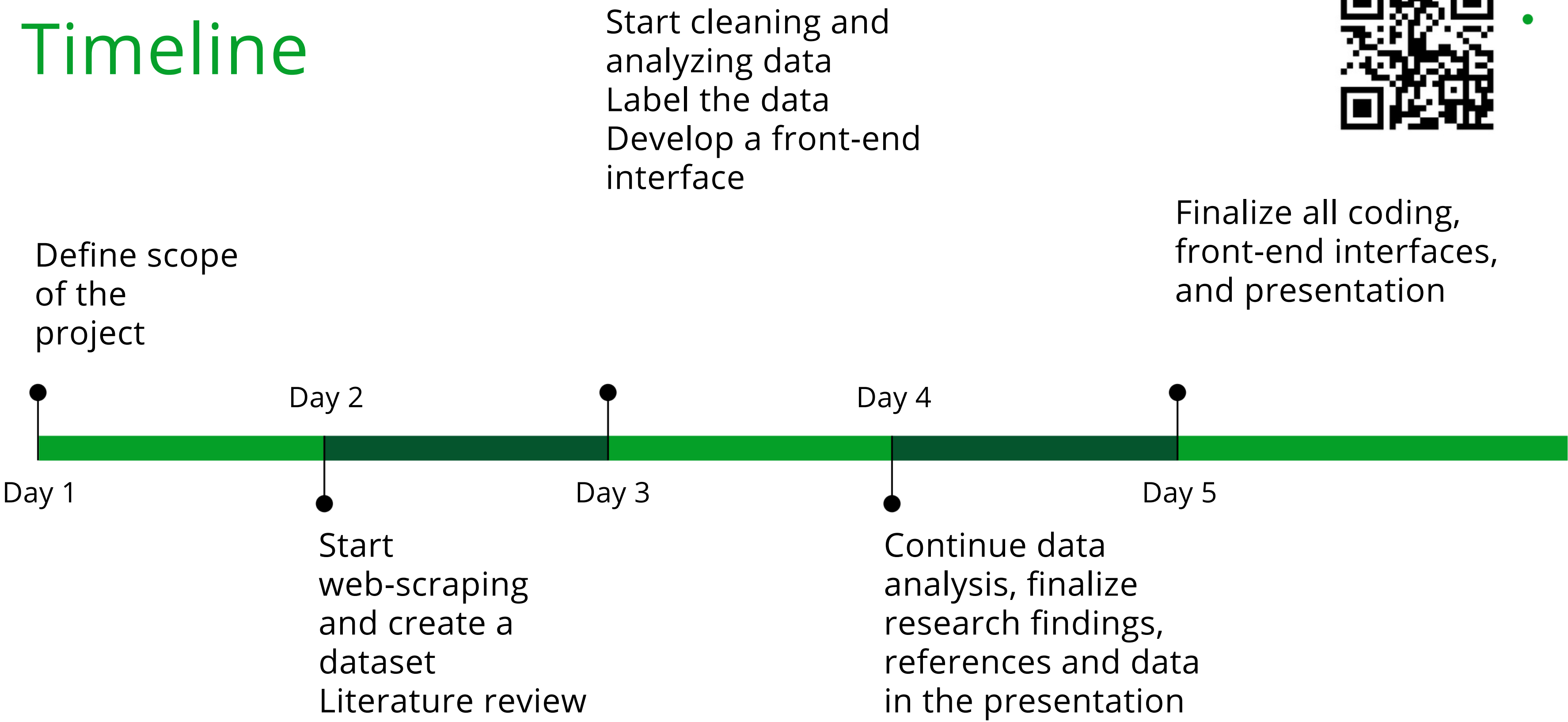
Note: This is a group project under DESN 9002 at the University of Hong Kong, completed in October 2023.

I was leading the data team. Specifically, I was in charge of team coordination, few-shot learning of GPT-3, incorporating medical Chatbot from Hugging Face for Question-Answer functions. I also oversaw the data pre-processing procedure and did part of the presentation.

CONTAG

Stay Informed. Stay Safe.

Timeline



Day 1

Define scope of the project

Day 2

Start web-scraping and create a dataset
Literature review

Day 3

Start cleaning and analyzing data
Label the data
Develop a front-end interface

Day 4

Continue data analysis, finalize research findings, references and data in the presentation

Day 5

Finalize all coding, front-end interfaces, and presentation

humanitarian crisis

Severity of the crisis

Number of people affected

Scale of disaster

Extent of emergency

Timeframe of addressing shock onset

Further aggravation of pre-existing vulnerabilities

Case Study

Contagious illnesses in India

Developing country

Lack of adequate healthcare infrastructure

Relatively high mortality rate & widespread health harm

Large English-speaking population

https://health.economictimes.indiatimes.com/news/industry/health-infrastructure-capacity-building-investments-the-focal-points-to-improve-healthcare-in-india/101970232

How many people in India speak English. Investor Times. (2023, September 6). https://investortimes.com/how-many-people-in-india-speak-english/

India: Who coronavirus disease (covid-19) dashboard with vaccination data. World Health Organization. (n.d.). https://covid19.who.int/region/searo/country/in

Defining an Outbreak

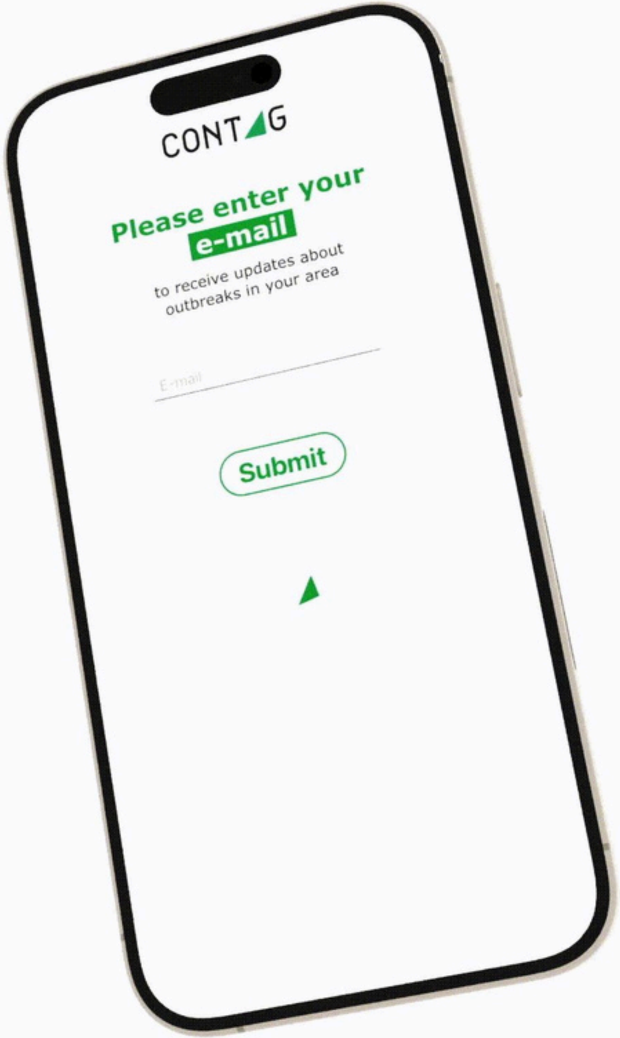
A sudden spike in the occurrence of a disease

- 🚑 Size of population at risk
- 🦠 Type of disease
- ⌛ Time period
- 💥 Effectiveness

➡️ **Mortality**

1,000 deaths per day

CONTAG



CONTAG

Please enter your **e-mail**

to receive updates about outbreaks in your area

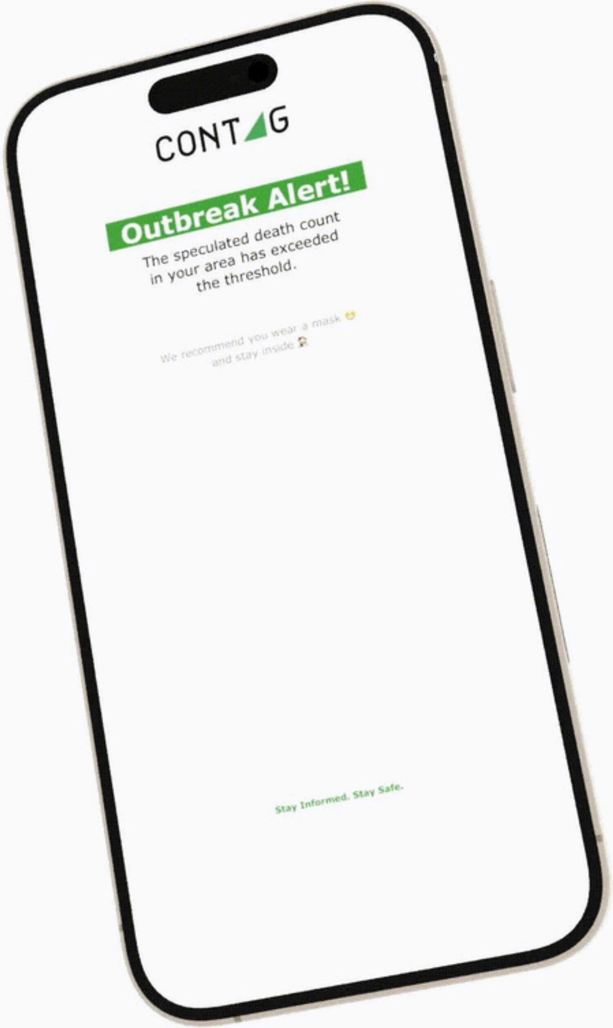
E-mail

ConTag Alert

Warning message

Short caption

Health advice



CONTAG

Outbreak Alert!

The speculated death count
in your area has exceeded
the threshold.

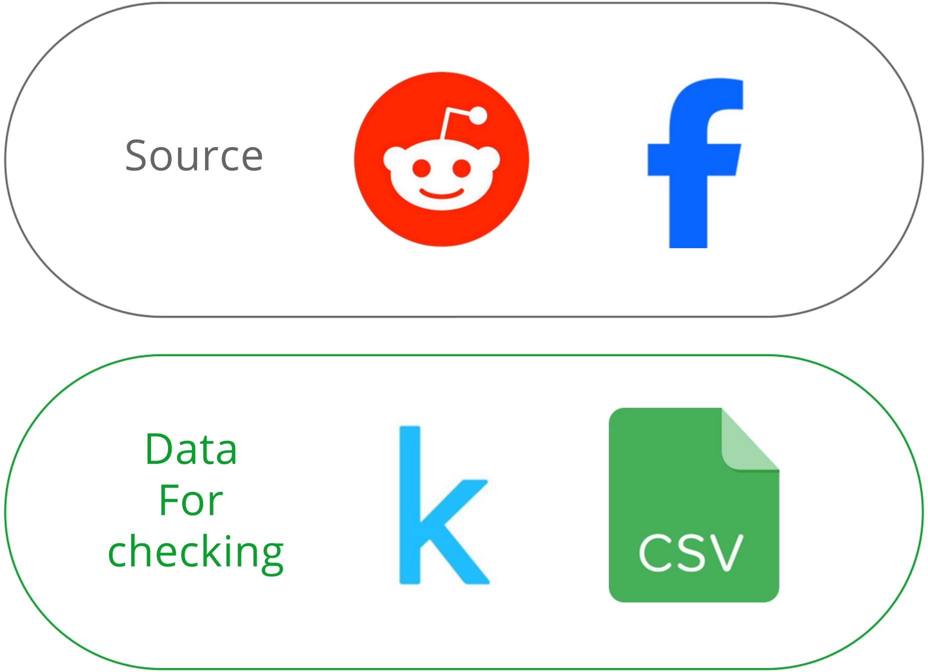
We recommend you wear a mask 😷
and stay inside 🏠

Stay Informed. Stay Safe.

Breakdown of Our Algorithm

1. Web Scraping
2. Data Cleansing
3. Labeling & Classification of Posts
4. Sentiment and severity analysis
5. Final Classification
6. Contag Alert

Web Scraping



Source

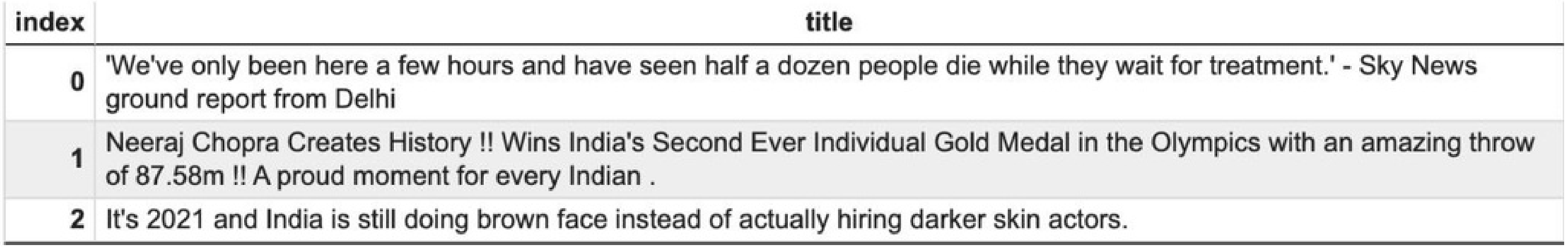
Data
For
checking

Data Cleansing

Objective: Assess the illness situation of posts

Title + Content + Hashtags = 1 data point

5825 data points in total

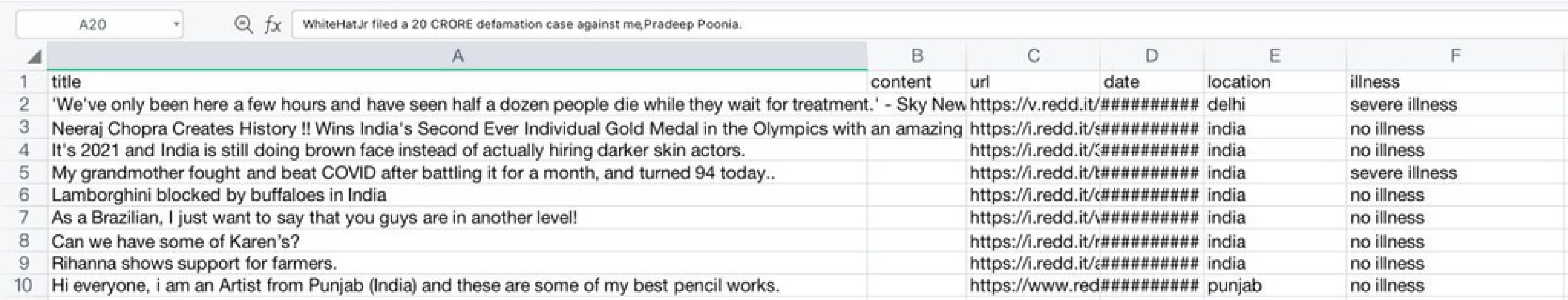


index	title
0	'We've only been here a few hours and have seen half a dozen people die while they wait for treatment.' - Sky News ground report from Delhi
1	Neeraj Chopra Creates History !! Wins India's Second Ever Individual Gold Medal in the Olympics with an amazing throw of 87.58m !! A proud moment for every Indian .
2	It's 2021 and India is still doing brown face instead of actually hiring darker skin actors.

Classification of posts

“What happens when labeled clean data isn’t available?”

Manual Labeling (few-shot learning)



A20 fx WhiteHatJr filed a 20 CRORE defamation case against me,Pradeep Poonia.

	A	B	C	D	E	F
1	title	content	url	date	location	illness
2	'We've only been here a few hours and have seen half a dozen people die while they wait for treatment.' - Sky New		https://v.redd.it/	##########	delhi	severe illness
3	Neeraj Chopra Creates History !! Wins India's Second Ever Individual Gold Medal in the Olympics with an amazing		https://i.redd.it/s	##########	india	no illness
4	It's 2021 and India is still doing brown face instead of actually hiring darker skin actors.		https://i.redd.it/3	##########	india	no illness
5	My grandmother fought and beat COVID after battling it for a month, and turned 94 today..		https://i.redd.it/t	##########	india	severe illness
6	Lamborghini blocked by buffaloes in India		https://i.redd.it/c	##########	india	no illness
7	As a Brazilian, I just want to say that you guys are in another level!		https://i.redd.it/v	##########	india	no illness
8	Can we have some of Karen's?		https://i.redd.it/r	##########	india	no illness
9	Rihanna shows support for farmers.		https://i.redd.it/a	##########	india	no illness
10	Hi everyone, i am an Artist from Punjab (India) and these are some of my best pencil works.		https://www.red	##########	punjab	no illness

OpenAI

ChatGPT API → A large amount of well-labeled data

LLM Labeling

Our machine learning models

The models are tuned using data from kaggle and data scraping
The models help collect data corresponding to the objective

Severity models (GPT 3.5 & 4)

Categorize each disease related post into severe, not severe, none

Disease model (Bert)

Classify disease and non-disease posts

Sentiment models (Bert)

Disease-related posts are paired with an emotion and measured by its level of positivity on a scale of 1 to 5

Why Sentimental Analysis?

Variability in Twitter Content Across the Stages of a Natural Disaster: Implications for Crisis Communication

Social media use rises during crisis

The public could give rapid and sustained attention on social media

Twitter transitions from being a medium of information providing to an outlet for affective responses and expressions of “fear”

Sci-Hub | Variability in Twitter Content Across the Stages of a Natural Disaster: Implications for Crisis Communication. Communication Quarterly, 63(2), 171–186 | 10.1080/01463373.2015.1012219. (2015). Sci-Hub.se. https://sci-hub.se/10.1080/01463373.2015.1012219

Sci-Hub | Visualizing Social Media Sentiment in Disaster Scenarios. Proceedings of the 24th International Conference on World Wide Web - WWW ’15 Companion | 10.1145/2740908.2741720. (2015). Sci-Hub.se. https://sci-hub.se/10.1145/2740908.2741720

The Final Classification Model

Comparing the database

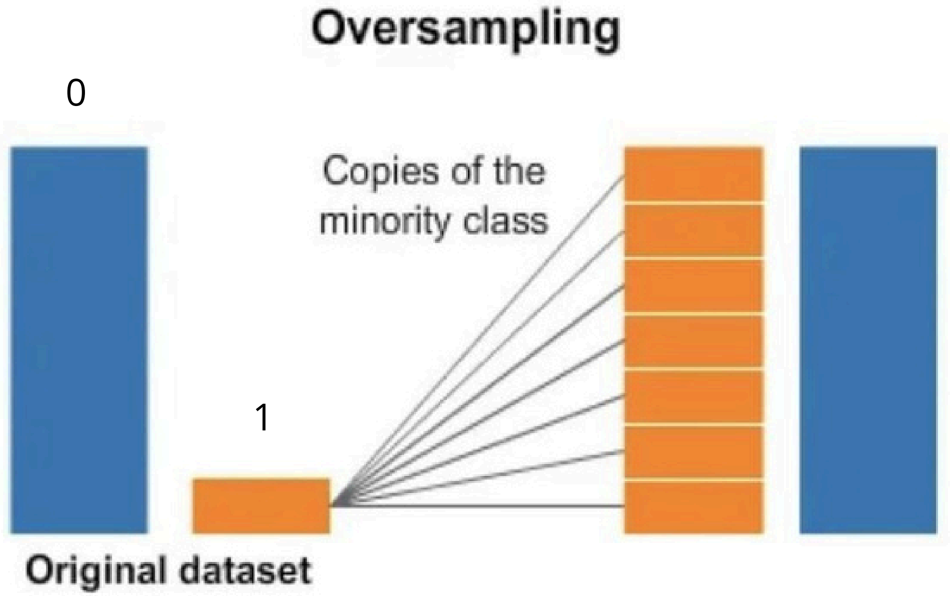
Used the WHO Dataset of COVID-19 deaths to determine when an outbreak occurred (>1000)

Features: average values obtained of posts on a day

Output: if there is a crisis

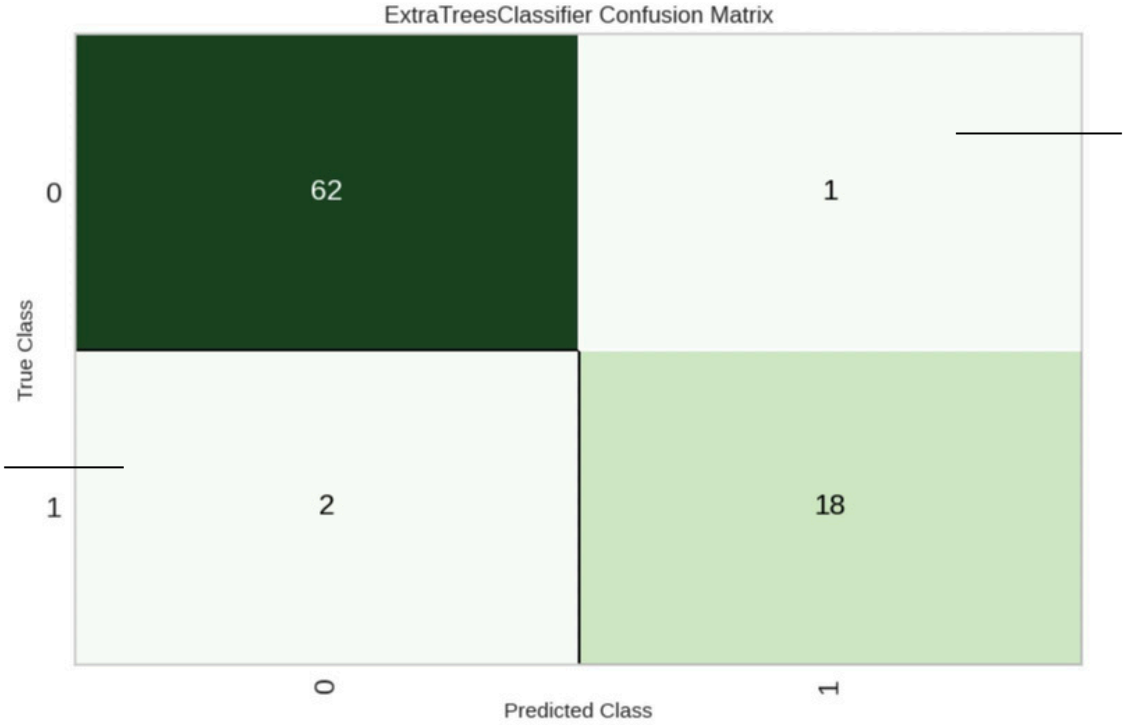
Tool: Pycarat : feature selection, training, tuning, stacking

Imbalanced data



Al-Serw, N. A.-R. (2021, February 21). Undersampling and oversampling: An old and a new approach. Medium. https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392

Final Output of the Classification Model



True Class \ Predicted Class	0	1
0	62	1
1	2	18

False Positive

False Negative

Ethical Considerations

Facts cannot be verified

Unable to tell if stories are trustworthy.

Subjective threshold

Cannot consider social, cultural, or political factors in defining an outbreak

Hinder website operations

Web scraping can impact users’ experiences on certain websites

Ethical Considerations

Privacy Concerns

Nonconsensual breach of personal information

Data Security

Data handling procedures and storage may be improved, intellectual property

Accessibility

The UX/UI is unable to cater to certain demographics

Improvements

Future/other diseases

Other locations

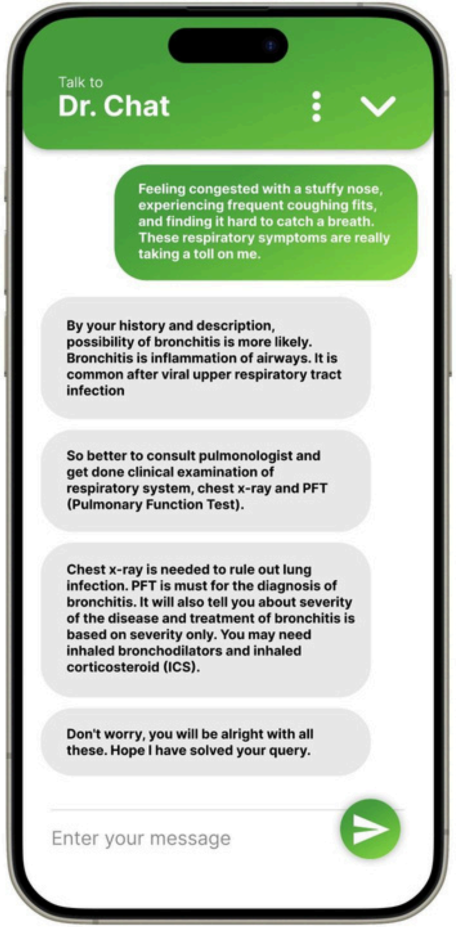
Site-specific outbreaks

Implementation across more social media sites

Accessibility for the partially impaired

Dr. Chat

Dr. Chat will intake your medical history and symptoms into consideration and provide medical advice or possibly a diagnosis



Feeling congested with a stuffy nose, experiencing frequent coughing fits, and finding it hard to catch a breath. These respiratory symptoms are really taking a toll on me.

By your history and description, possibility of bronchitis is more likely. Bronchitis is inflammation of airways. It is common after viral upper respiratory tract infection

So better to consult pulmonologist and get done clinical examination of respiratory system, chest x-ray and PFT (Pulmonary Function Test).

Chest x-ray is needed to rule out lung infection. PFT is must for the diagnosis of bronchitis. It will also tell you about severity of the disease and treatment of bronchitis is based on severity only. You may need inhaled bronchodilators and inhaled corticosteroid (ICS).

Don't worry, you will be alright with all these. Hope I have solved your query.

References

Health Infrastructure, Capacity Building & Investments: The focal points to improve healthcare in India - etETHealthworld. (2023, July 20). https://health.economictimes.indiatimes.com/news/industry/health-infrastructure-capacity-building-investments-the-focal-points-to-improve-healthcare-in-india/101970232

How many people in India speak English Investor Times. (2023, September 6). https://investortimes.com/how-many-people-in-india-speak-english/

India: Who coronavirus disease (covid-19) dashboard with vaccination data. World Health Organization. (n.d.). https://covid19.who.int/region/searo/country/in

Sci-Hub | Variability in Twitter Content Across the Stages of a Natural Disaster: Implications for Crisis Communication. Communication Quarterly, 63(2), 171–186 | 10.1080/01463373.2015.1012219. (2015). Sci-Hub.se. https://sci-hub.se/10.1080/01463373.2015.1012219

Sci-Hub | Visualizing Social Media Sentiment in Disaster Scenarios. Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion | 10.1145/2740908.2741720. (2015). Sci-Hub.se. https://sci-hub.se/10.1145/2740908.2741720

Al-Serw, N. A.-R. (2021, February 21). Undersampling and oversampling: An old and a new approach. Medium. https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392

What are the machine learning models we have in our algorithms?

1. Disease models

-objective: classify disease and non-disease posts

-we tune our disease model with reference from bert (pre-trained learning model)

-with the tuned model, we could apply more data sets to it, and collect more disease posts on social media, used in the following steps.

2. Sentiment model

-objective:to pair an emotion for each disease-related post; and to measure the level of positivity in the post with the scale of 1 to 5 -using data from kaggle/data scraping, we tune our sentiment model

-with the tuned model, we could collect datas corresponding to the objective, where we could utilize in the future process

3. Severity of Illness model

-objective: dividing the illness of a disease-related post into a)severe, b)not severe, c)none

-using the same concept as above with reference to gpt 3.5/4, we could create desired data