CITATION

Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction. [Submitted]. *Bull World Health Organ*. E-pub: 19 March 2020. doi: http://dx.doi.org/10.2471/BLT.20.255695

CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction

CoronaTracker Community Research Group

Fairoza Amira Binti Hamzah[a], Cher Han Lau[b], Hafeez Nazri[c], Dominic Vincent Ligot[d], Guanhua Lee[e], Cheng Liang Tan[f], Mohammad Khursani Bin Mohd Shaib[g], Ummi Hasanah Binti Zaidon[h], Adina Binti Abdullah[i], Ming Hong Chung[j], Chin Hwee Ong[k], Pei Ying Chew[l] and Roland Emmanuel Salunga[m]

[a] The Kyoto College of Graduate Studies for Informatics
[b] Lead
[c] Media Prima Digital Sdn Bhd
[d] Cirrolytix
[e] National University Health System
[f] International Medical University
[g] Vase.ai
[h] Institut Wanita Berdaya Selangor
[i] University of Malaya
[j] Quest International University Perak
[k] ST Engineering
[l] Universiti Malaysia Sarawak
[m] National Capital Region, Philippines

Correspondence to Fairoza Amira Binti Hamzah (email: fairozaamira@gmail.com)
(Submitted: 18 March 2020 – Published online: 19 March 2020)

CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction

CoronaTracker Community Research Group

Abstract

Background

COVID-19 outbreak was first reported in Wuhan, China and has spread to more than 50 countries. WHO declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on 30 January 2020. Naturally, a rising infectious disease involves fast spreading, endangering the health of large numbers of people, and thus requires immediate actions to prevent the disease at the community level. Therefore, CoronaTracker was born as the online platform that provides latest and reliable news development, as well as statistics and analysis on COVID-19. This paper is done by the research team in the CoronaTracker community and aims to predict and forecast COVID-19 cases, deaths, and recoveries through predictive modelling. The model helps to interpret patterns of public sentiment on disseminating related health information, and assess political and economic influence of the spread of the virus.

Methods:

Real-time data query is done and visualized in our website, then the queried data is used for Susceptible-Exposed-Infectious-Recovered (SEIR) predictive modelling. We utilize SEIR modelling to forecast COVID-19 outbreak within and outside of China based on daily observations. We also analyze the queried news, and classify the news into negative and positive sentiments, to understand the influence of the news to people's behavior both politically and economically.

Findings:

At the time of writing this paper, the number of confirmed cases is expected to exceed 76000 cases,

and reach the peak of this outbreak before 20 February 2020. The average Infected-Suspected ratio was found to be 2.399 which we used to initialize the number of Exposed individuals as a product of the number of Infected individuals on 20 Jan 2020. This outbreak is assumed to reach its peak in late May 2020 and will start to drop around early July 2020. Based on the news queried in our system, we found that there are more negative articles than positive articles, and displayed similar words for both negative and positive sentiments. The top five positive articles are about collaboration and strength of individuals in facing this epidemic, and the top five negative articles are related to uncertainty and poor outcome of the disease such as deaths.

Conclusions:

COVID-19 is still an unclear infectious disease, which means we can only obtain an accurate SEIR prediction after the outbreak ends. The outbreak spreads are largely influenced by each country's policy and social responsibility. As data transparency is crucial inside the government, it is also our responsibility not to spread unverified news and to remain calm in this situation. The CoronaTracker project has shown the importance of information dissemination that can help in improving response time, and help planning in advance to help reduce risk. Further studies need to be done to help contain the outbreak as soon as possible.

Keywords: COVID-19, data analysis, sentiment analysis, predictive modelling, SEIR

1. Introduction

On 31 December 2019, the first reported case in the COVID-19 outbreak was reported in Wuhan, China. The first case outside of China was reported in Thailand on 13 January 2020 [1]. Since then, this ongoing outbreak has now spread to more than 50 other countries [2]. WHO declares COVID-19 outbreak as a Public Health Emergency of International Concern (PHEIC) by WHO on 30 January 2020 [3]. There are over 76,000 cases of confirmed COVID-19 worldwide as

of 20 February [2].

An infectious disease outbreak is the occurrence of a disease that is not usually expected in a particular community, geographical region, or time period [4]. Typically, a rising infectious disease involves fast spreading, endangering the health of large numbers of people, and thus requires immediate action to prevent the disease at the community level [5]. COVID-19 is caused by a new type of coronavirus which was previously named 2019-nCoV by the World Health Organization (WHO). It is the seventh member of the coronavirus family, together with MERS-nCoV and SARS-nCoV, that can spread to humans [1]. The symptoms of the infection include fever, cough, shortness of breath, and diarrhea. In more severe cases, COVID-19 can cause pneumonia and even death [6]. The incubation period of COVID-19 can last for 2 weeks or longer [7]. During the period of latent infection, the disease may still be infectious. The virus can spread from person to person through respiratory droplets and close contact [8].

An 'infodemic' has accompanied the COVID-19 outbreak which is essentially an over-abundance of information regarding the outbreak. As some of the information available to the public may not be accurate, it becomes hard for people to find reliable sources and trustworthy guidance when they need it [9]. Because of the high demand for appropriate and trustworthy information about 2019-nCoV, WHO technical risk communication and social media teams have been working closely to track and respond to myths and rumors via its headquarters in Geneva, its six regional offices and its partners. The organization is working continuously to identify the most widespread rumors that can possibly harm the public's health, such as inaccurate prevention measures or claims of cures. These myths are then rebutted with evidence-based information. WHO is making public health information and advice on the COVID-19, including myth busters, accessible on its social media channels (including Weibo, Twitter, Facebook, Instagram, LinkedIn, Pinterest) and on their website [10].

Communication during emerging pandemics presents a distinctive public health education task. Health consumers must be informed about an impending health threat [11]. However, there may be difficulties in providing accurate information regarding the outbreak in the initial stage. This is mainly related to the high degree of uncertainty about the exact route of transmission, treatment of the infections, and prospects of recovery in an outbreak. All countries need to prepare existing public health communication networks, media and community engagement staff for a possible case in their country, as well as for the appropriate response if it happens. The governments should coordinate communications with other response organizations and include the community in response operations. WHO stands ready to coordinate with partners to support countries in their communication and response to community engagement.

To ensure a people-centered response to COVID-19, an expanding group of global response organizations such as the United Nations Children's Fund (UNICEF) and the International Federation of Red Cross and Red Crescent Societies (IFRC) are coordinating efforts with WHO to apply biomedical recommendations at the community level. These organizations are active at the global, regional and country level to ensure that affected populations have a voice and are part of the response. Ensuring that global recommendations and communication are tested and adapted to local contexts will help countries to gain better control over the COVID-19 outbreak [10].

Peoples' response to the news about a spreading contagious disease is likely to lead to increased anxiety and amplification of risk perceptions [11]. Social media networks have functioned as channels for firsthand information from which the public can acquire disease-related information during infectious disease outbreaks. These platforms also enable simple and quick sharing of information with family, friends, and neighbors in real time [12]. For example, the Ministry of Health in Malaysia have been uploading posts related to COVID-19 to educate the public since 19 January [13] and their Director General of Health is also active on his own

Facebook page to clear confusion and doubts for the public [14]. This is important when traditional forms of media are unable to provide relevant and timely information to the public. Social media now serves as a major, immediate information source but while the focus of latest information has been on the role of social media during infectious disease outbreaks, the question that should be brought to light is still, how the use of social media may trigger the public's emotional or noncognitive response, affect perception of risk, and preventive behaviors [10].

Therefore, CoronaTracker [15] was born as the online platform that provides the latest and verified news development, statistics and analysis on COVID-19. This platform is a community-based project initiated on 25th February out of concerns on the outbreak that halted Mainland Chinese of Lunar New Year's celebration. The CoronaTracker website was launched on 27th January 2020, after two days working relentlessly, and has gathered more than 1300 volunteers across the globe. This paper is a part of a work by the research team of CoronaTracker community. The main objective of this paper is to predict and forecast COVID-19 cases, deaths, and recoveries through predictive modelling, and to decipher patterns on public sentiment related to health information dissemination. At the same time, assess the political and economic impact of the virus spread.

We propose a comprehensive framework to manage health information data as a tool for public health practitioners in managing epidemics and crafting public health response and policy. This study focuses on the role of audiences in the process of disseminating health risk information and examines behaviors that contribute to information amplification upon hearing the news.

The structure of this paper is as follows; Section 1 introduces COVID-19 and CoronaTracker community, as well as explains the significance of this research. Section 2 describes on related works in predictive modelling of the paper and news-based sentiment analysis for this research on psychological, politics and economics aspects. Section 3 explains our study design and

methodologies. Section 4 presents our findings in current trends, predictive modelling and sentiment analysis of the outbreak. Our findings are discussed in Section 5 and this paper is concluded in Section 6.

2. Related Works

SEIR refers to Susceptible, Exposed, Infectious, and Removed or Recovered, respectively. It is based on the SIR model but adds the Exposed compartment as a variable. Susceptible refers to individuals who can catch the infection and may become hosts if exposed, Exposed are individuals who are already infected but are asymptomatic, Infectious are individuals who are showing signs of infection and can transmit the virus, Removed or Recovered are individuals who are previously infected but are no longer infectious and already immune to the virus [16].

Once the compartments of SIR or SEIR models are determined, modelling can be done using a variety of methods. In [17], a Conditional Autoregressive (CAR) was used to account for epidemics with a spatial or transportation-related vector and modelled with MCMC. In [16], demographic effects such as birth and death rates were added to the SEIR to model equilibria with vital dynamics.

Sentiment analysis is a supervised machine learning problem. There are different types of sentiment analysis including fine-grained sentiment analysis, emotion detection, aspect-based sentiment analysis and multilingual sentiment analysis. In binary sentiment classification, the possible categories are positive and negative. In fine-grained sentiment classification, there are five groups (very negative, negative, neutral, positive, and very positive). Sentiment analysis is one of the most popular tasks in natural language processing, and there has been a lot of research and progress in solving this task accurately [18].

Deep neural networks are widely used in sentiment polarity classification; however, it often requires huge numbers of training data, and the size of training data varies quite significantly

among domains. In [19], it was found that a dual-module approach is the best method that encourages the learning of models with promising generalization abilities. Bidirectional Encoder Representations from Transformers (BERT) is an embedding layer designed to train deep bidirectional representations from unlabeled texts by jointly conditioning on both left and right context in all layers. It is pretrained from a large unsupervised text corpus such as Wikipedia or BookCorpus. There are 15% of the words in the input sequence are masked out which is one of the objectives of BERT. Then, a deep bidirectional Transformer encoder is fed by the entire sequences so that the model learns to predict the masked words.

Moreover, this small model has been trained on SST-2 dataset which is a common dataset for sentiment-analysis [20]. However, there are few disadvantages in this method as it is based on SST-2 dataset which is for movie reviews and our dataset is about coronavirus news. It is a similar task which is for sentiment analysis but it does not perform that well because sentiment for movies and news might be different. However, it is the fastest way to get results and act as a benchmark or starter for further research. It can also be easily improved by adding more dataset for our domain (coronavirus news). Last but not least, it can do prediction instantly compared to previous methods that need bigger computer resources.

3. Methods

3.1 Data Source

Data is extracted from verified sources such as John Hopkins University [21], WHO and DingXiangYuan, a website authorized by the Chinese government. The sites reported confirmed COVID-19 cases, as well as recovered and deaths for affected countries and regions. Details on how our team fetched the data is in Section 3.2.

3.2 Data Visualization

The data collected in CoronaTracker is available on data lakes platform. Both the

platform and dashboard are hosted in Amazon Web Services (AWS). We provisioned AWS Relational Database Service (RDS) to host the data in MySQL table form. All the data collected and ingested using Python program running in AWS Elastic Compute Cloud (EC2) and was scheduled to automatically update every 15 minutes. The size of the database is relatively around 30GB.

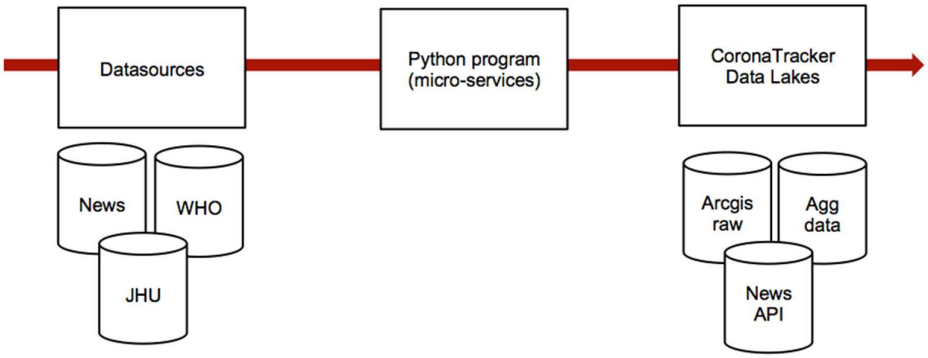


Fig. 1 High level diagram ingestion for CoronaTracker

We developed our own micro-services and scraper in Python to fetch the data and news from the sources. Python has been the best open source platform to use because of the less level of complexity, readily libraries and easy to deploy in production environment (EC2). The code snippet for our data fetcher is shown as in Fig. 2.

```
Input:  A json fetcher from API url from John Hopkin University data API
Output: A structured data in SQL table (arcgis table) which will be auto-
fetched data every 15 minutes
--------------------------------------------------------------------------
u = 'JHU data API url'
r=requests.get(u)
e=sa.create_engine('mysql://dsn')
x=json.loads(r.content)

d=dd[['attributes.Province_State', 'attributes.Country_Region',
'attributes.Last_Update',
    'attributes.Lat', 'attributes.Long_', 'attributes.Confirmed',
'attributes.Deaths','attributes.Recovered' ]]
```

Fig. 2 Code snippet for data fetcher

The fetched data will be stored in relational database, MySQL. The micro-services will crawl the data every 15 minutes and store it in table form. The data mostly will be stored in raw format before being aggregated for visualization. Aggregate functions are built using SQL statement like below snippet. Aggregation is important to show latest of the sum values for every country in the table.

```
SELECT
  CAST(SUM(confirmed) AS UNSIGNED) AS confirmed,
  CAST(SUM(deaths) AS UNSIGNED) AS deaths,
  CAST(SUM(recovered) AS UNSIGNED) AS recovered,
  MAX(posted_date) as created
FROM
  arcgis
WHERE
  posted_date = (SELECT MAX(posted_date) FROM arcgis)
LIMIT 1
```

Fig. 3 SQL statement and AGG (aggregate) functions

Example of the data in the databases is shown in Fig. 4. We also geo-coded the location for easier mapping during visualization.

	123 agg_confirmed	123 agg_death	123 agg_recover	agg_date
21	60,441	1,370	6,280	2020-02-13
22	64,541	1,384	7,171	2020-02-14
23	67,178	1,527	8,578	2020-02-15
24	69,337	1,669	9,624	2020-02-16
25	72,338	1,775	11,396	2020-02-17
26	73,773	1,875	13,124	2020-02-18
27	75,905	2,012	15,084	2020-02-19
28	76,389	2,130	16,882	2020-02-20
29	77,424	2,248	18,864	2020-02-21
30	78,560	2,362	21,259	2020-02-22
31	79,568	2,466	23,386	2020-02-23
32	80,826	2,699	27,683	2020-02-25
33	81,922	2,770	30,311	2020-02-26
34	83,226	2,810	33,252	2020-02-27
35	84,540	2,867	36,686	2020-02-28
36	86,361	2,933	39,761	2020-02-29
37	88,123	2,990	42,670	2020-03-01
38	89,948	3,050	45,394	2020-03-02
39	93,005	3,134	48,192	2020-03-03
40	94,955	3,216	51,041	2020-03-04
41	96,887	3,305	53,638	2020-03-05
42	100,312	3,408	55,690	2020-03-06

Fig. 4 Result from AGG query

For every 15 min, the cumulative case counts will be updated for all provinces and other affected countries. In the beginning, we found that WHO and JHU data are relatively slow compared to other sources, thus we implement manual update to our system after verifications to allow more real-time data.

3.3 Predictive Modelling – SEIR Model

Here we will briefly discuss the properties of basic Susceptible-Exposed-Infected-Removed (SEIR) system that will be used to describe the recent outbreak of COVID-19 in China [22]. We considered a simple SEIR epidemic model for the simulation of the infectious-disease spread. Individuals were each assigned to one of the following disease states; Susceptible (S), Exposed (E), Infections (I) and Removed (R) which refers to segment not yet infected and disease-free, individuals that are experiencing incubation duration, the confirmed (isolated) cases, recovered individuals, respectively. The SEIR diagram in Fig. 5 shows how individuals move

through each compartment in the model.

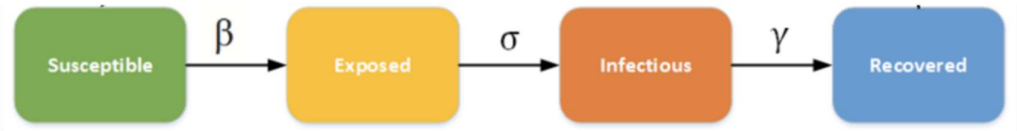


Fig. 5 SEIR model with four states **[23]**

Parameters within this model are:

1. β, controls the rate of spread, which represents the probability of transmitting disease between a susceptible and an infectious individual [24]. The reproductive number used in this paper is 2.2.
2. Incubation rate σ, is the rate of latent individuals becoming infectious. Given the known average duration of incubation Y, $\sigma = 1/Y$ [24]. The average incubation duration of 5.2 days are used here.
3. Recovery rate $\gamma = 1/D$, is determined by the average duration of recovery D, of infection. After this period, they enter the removed phase. The average duration of infection is calculated as the average serial interval minus the average incubation duration. The average serial interval of 7.5 days is used in this paper [24]. 2.3 days of an average infectious duration is used here.

Fig. 6 shows the diagrammatic representation of virus progress in an individual, where infectious occurs at t_L, during the latent period, infected individual is not infectious, and at t_{sy}, symptoms appear [16]. The first transmission to the left healthy individuals is at t_{tr}. After t_R, the removed (recovered) people are considered no longer infectious.

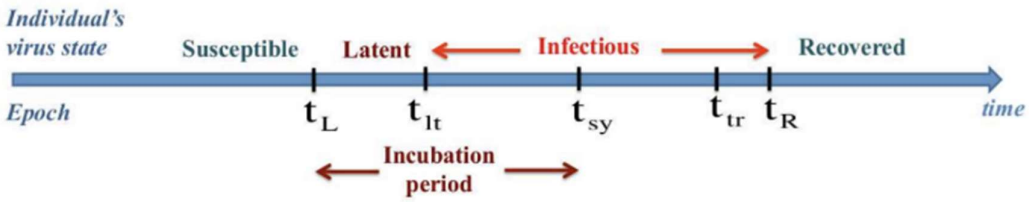


Fig. 6 Virus progress in an individual by using the SEIR model

We can describe the virus transmission by the following nonlinear ordinary differential equation [23] as shown in equation (10) to (13).

$$\frac{dS}{dt} = -\frac{\beta SI}{N}, \qquad (10)$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E, \qquad (11)$$

$$\frac{dI}{dt} = \sigma E - \gamma I, \qquad (12)$$

$$\frac{dR}{dt} = \gamma I, \qquad (13)$$

3.4 Sentiment Analysis

After we have stored the news inside the CoronaTracker database, we extract news description as it contains a summary of the news that is neither too short nor too long, which can be bad for the model we are going to use otherwise. We only select descriptions that are at least more than 8 words, and discard non-English descriptions because the pre-trained model we use have been trained on SST-2 [25], which is a dataset for sentiment analysis for English language. We use a library called transformers by huggingface [26]. The input sentences will be separated by their respective polarity for further analysis like topic modelling and generating word cloud for each polarity.

4. Findings

4.1 Current Outbreak Trends

Fig. 7 to Fig. 9 shows the current trends for COVID-19 outbreak as displayed in CoronaTracker website [27]. The cases reported are visualized in analytics dashboard to show the outbreak trend for confirmed, recovered and deaths cases for all regions and countries. This aligns with our objectives to show the outbreak progress over the period of time for each segment. It was found that the total number of confirmed cases for all countries and regions are increasing steadily, but on day 24, the huge increment with 15000 differences were discovered. This is due to the change

in how China measured the confirmed cases. Fig. 9 is plotted by using an open source mapping library from Leaflet [28]. Table 1 shows the details on reported cases according to the countries [27]. Both are retrieved on 3rd March 2020.

Fig. 7 Outbreak trend over time

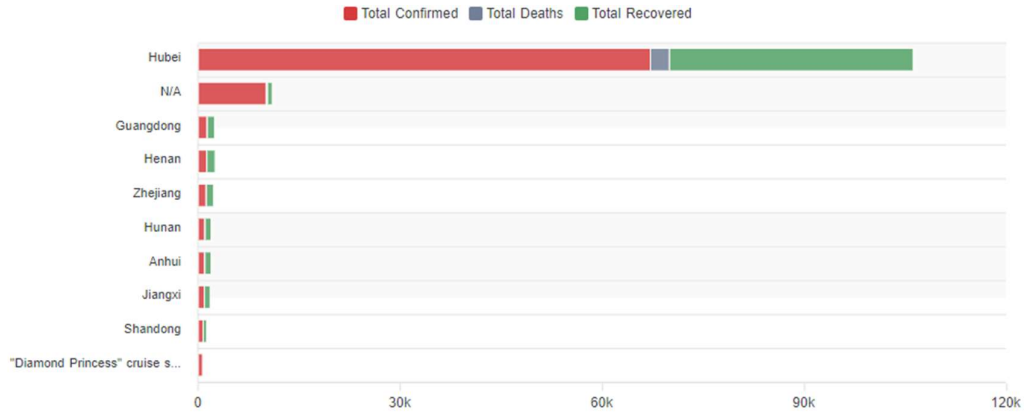


Fig. 8 Most affected regions

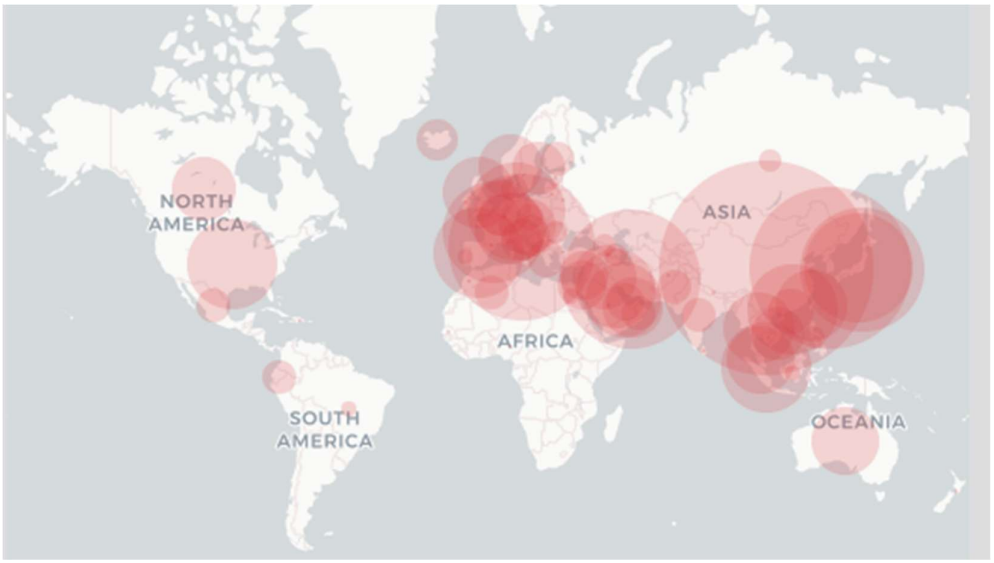


Fig. 9 World map of affected region, where the darker the redness in the map, the more the number of infected cases detected.

Table 1 Reported cases as of 3rd March 2020

Country	Total Confirmed	Total Recovered	Total Deaths
China	80,151	47,367	2,945
South Korea	5,186	30	28
Italy	2,036	149	52
Iran	1,501	291	66
Others*	706	10	6
Japan	274	32	6
France	191	12	3
Germany	165	16	0
Spain	120	2	0
Singapore	108	78	0
United States	105	7	6

Hong Kong	100	37	2
Kuwait	56	0	0
Bahrain	49	0	0
Thailand	43	31	1
Taiwan	42	12	1
Switzerland	42	0	0
United Kingdom	40	8	0
Malaysia	36	22	0
Australia	31	11	1
Canada	27	6	0
Iraq	26	0	0
Norway	25	0	0
United Arab Emirates	21	5	0
Netherlands	18	0	0
Austria	18	0	0
Vietnam	16	16	0
Sweden	15	0	0
Lebanon	13	0	0
Israel	12	1	0
Macau	10	9	0
Iceland	9	0	0
Croatia	9	0	0
Belgium	8	1	0
San Marino	8	0	0

Greece	7	0	0
Qatar	7	0	0
Finland	6	1	0
Oman	6	1	0
Ecuador	6	0	0
India	5	3	0
Mexico	5	1	0
Algeria	5	0	0
Pakistan	5	0	0
Denmark	4	0	0
Czech Republic	4	0	0
Russia	3	2	0
Philippines	3	2	1
Georgia	3	0	0
Romania	3	0	0
Azerbaijan	3	0	0
Egypt	2	1	0
Brazil	2	0	0
Portugal	2	0	0
Indonesia	2	0	0
Cambodia	1	1	0
Nepal	1	1	0
Sri Lanka	1	1	0
Afghanistan	1	0	0

Estonia	1	0	0
Andorra	1	0	0
Nigeria	1	0	0
Luxembourg	1	0	0
Saudi Arabia	1	0	0

* Cases identified on a cruise ship currently in Japanese territorial waters.

4.2 Predictive Modelling

In this section, we will model the global trajectory of the infection counts using the SEIR model, 240 days from the start date of 20 January 2020. This start date was chosen because earlier dates were assumed as "burn-in" for the reporting of infection counts. The parameters of the SEIR model were determined in section 3.2 and the world's population are assumed to be 7.5 billion people.

Using the data source from JHU, we lacked an initial number of Exposed individuals. To gather this information, we made the estimation that the number of Infected individuals today is an approximate ratio to the number of suspected individuals 6 days ago with an incubation duration of 5.2 days; we rounded up the incubation duration from 5.2 days for this calculation. The data to calculate this ratio was collected via real-time query from DingXiangYuan which provided suspected and infected counts in mainland China. Using two weeks of data from 20 Jan 2020, the average Infected-Suspected ratio was found to be 2.399 which we used to initialize the number of Exposed individuals as a product of the number of Infected individuals on 20 Jan 2020.

Using Scipy's implementation [29] of the numerical integration of ordinary differential equations, *odeint*, we plotted the E and I trajectories of the SEIR compartment. The Exposed and Infected trajectories tell us the number of individuals in those compartments over time. It can be seen from the plot that the maximum number of Infected individuals is 425.066 million globally

on 23 May 2020. Thereafter, the number of Infected individuals dropped to under 10 million on 12 July 2020, under 1 million on 3 Aug 2020 and under 10,000 at the end of the trajectory on 14 Sep 2020.

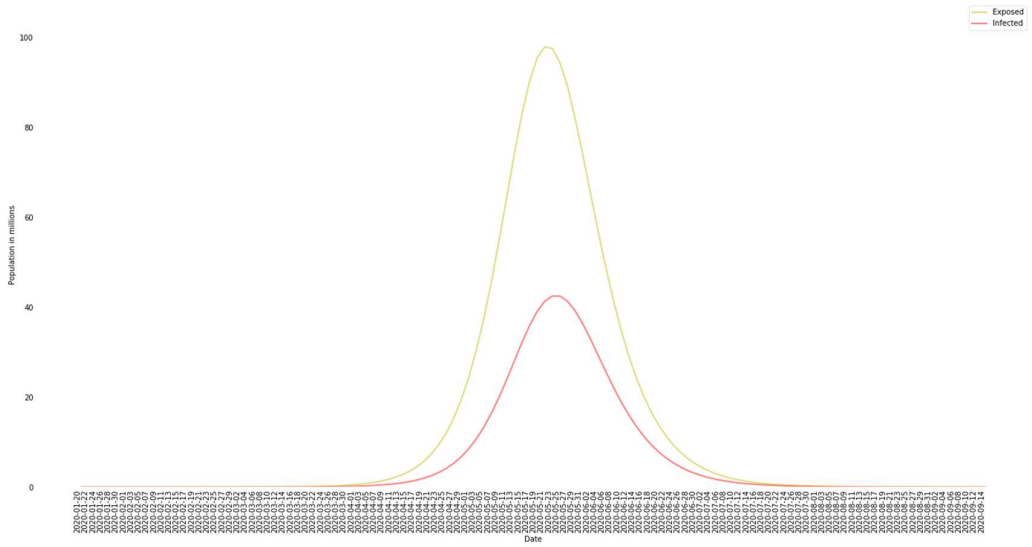


Fig. 10 Simulated Exposed and Infected Per Day

Although the SEIR model is a numerical simulation, the numbers provide us a degree to which the COVID-19 cases can surge to. These trajectories could serve as a means for governments, businesses and individuals to plan and mitigate for such a spike in Infected cases. Everyone should work towards blunting the curve and stopping the spread as per instructions set out by their governments on personal hygiene, control measures and refraining from mass gatherings.

4.3 Sentiment Analysis

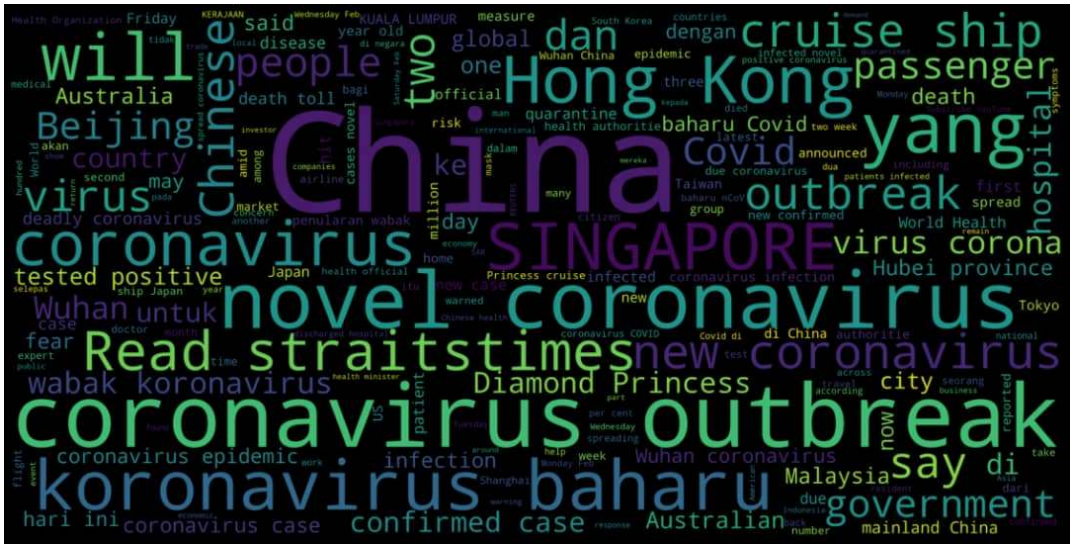
This paper also presents a sentiment analysis of recent verified news to understand peoples' reaction psychologically, politically and economically. Table 2 shows the number of positive and negative articles analysed from the news.

Table 2 Number of positive and negative articles from the news

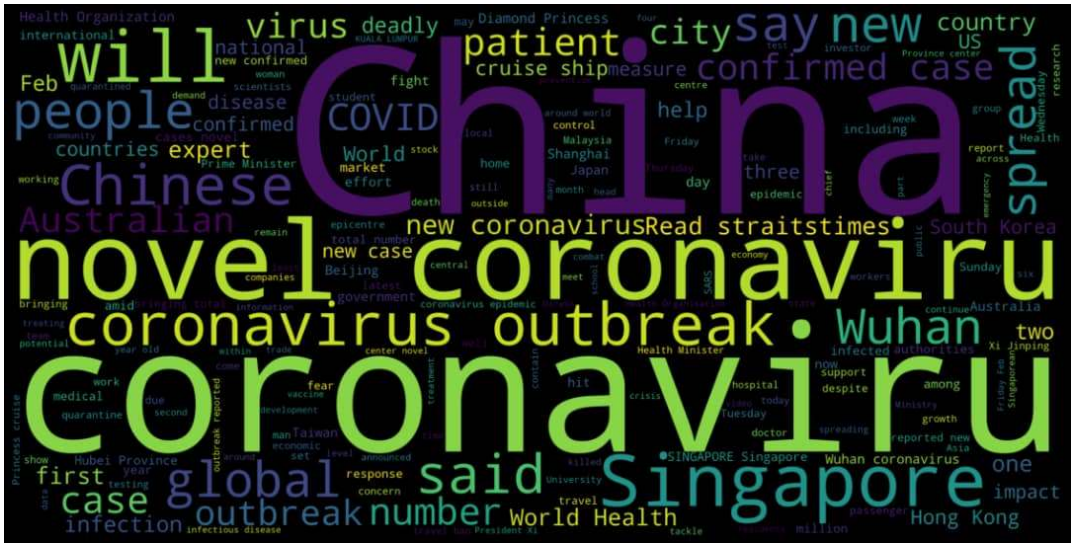
Sentiment	n	Sentiment	n

Positive	561	Negative	2548

Fig. 11(a) and (b) show the word cloud for positive and negative sentiments, respectively. It was found that both sentiments have the word “China” but it is necessary to check the word sequences to identify whether sentences that contain the word “China” are considered positive or negative.



(a) Positive sentiments



(b) Negative sentiments

Fig. 11 Word cloud for sentiments analysis on the news

Top 5 positive and negative articles are represented in Table 3 and 4.

Table 3 Top 5 positive articles

News	No. of words	Sentiment type	Score
MELBOURNE: The coach of the Chinese women's soccer team has praised the "strong hearts" of his players after they emerged undefeated from an ...	24	POSITIVE	0.999845
WHO Director-General said solid collaboration, transparency, prompt sharing of data, and accurate ad	13	POSITIVE	0.999836
Chinese President Xi Jinping has written a letter expressing thanks to the Bill & Melinda Gates Foundation for the organisation's "generosity" and ...	23	POSITIVE	0.999809
SINGAPORE: The reaction of Singaporeans, as they pitch in to help with the Wuhan coronavirus threat, has been 'amazing', said Law and Home Affairs Minister K Shanmugam on Sunday (Feb 2).	31	POSITIVE	0.999795
"We very much appreciate the efforts of the doctors. They took good care of us," said a patient surnamed Li while walking toward her	35	POSITIVE	0.999698

family after bidding farewell to the medical workers in Hongshan Gymnasium.

Table 4 Top 5 Negative Sentiments Articles

News	No. of words	Sentiment type	Score
SINGAPORE: The Wuhan coronavirus will cause current economic uncertainties to escalate, but Singapore's economy is prepared to weather the impending financial impact, said Manpower Minister Josephine Teo on Thursday (Jan 30).	31	NEGATIVE	0.966091
BEIJING - A Chinese teenage boy suffering from cerebral palsy was found dead at home in Hubei province six days after his father and younger brother were quarantined for suspected infection with a novel coronavirus.. Read more at straitstimes.com.	39	NEGATIVE	0.936342
An image of a Chinese man collapsed in a Seoul subway station, apparently too sick to stand, has been held up by many Koreans as proof of just how great a threat the	33	NEGATIVE	0.99643

It is not necessary to wear a face mask for now since no human-to-human transmission of the novel coronavirus (2019-nCoV) has occurred in Malaysia, says infe...	26	NEGATIVE	0.771767
US researchers are already working on a vaccine against the new virus that has emerged in China.	17	NEGATIVE	0.94461

5. Discussion

5.1 Health information dissemination

Information from news articles played an important role in empowering citizens during an epidemic. However, our analysis found more negative articles than positive articles which is a concern. (2548 vs. 561) The word clouds corresponding to the negative and positive statements displayed similar words. A content analysis of text mined Ebola on news articles and scientific publications showed similar findings. According to An's study, they found a difference in coverage of topics but a word co-occurrence map shared similar entities, which showed the limitation of single word tagging for sentiment analysis during an epidemic. An interface for citizens to record their sentiments and expressing their opinions on the news articles could contribute towards improving these findings [30]. The top 5 positive articles were about collaboration and strength of individuals in facing this epidemic, whereas the top 5 negative articles were related to uncertainty and poor outcome of disease like death. Another study evaluating information shared during the Ebola epidemic also found news articles to place more focus on event-related entities such as person, organization and location, while social media information like in Twitter places

covers more time-oriented entities [31].

5.2 Economic and politics impact

COVID-19 is spreading with astounding speed and have severe consequences. The cases increase rapidly with major outbreaks in South Korea, Italy, Iran, the United States and more than 50 other countries. Governments are put to the test with escalating high-pressure and costs of the outbreak in terms of public trust and the economy. Any mismanagement could carry political costs, as their legitimacy and competency will be called into question by the people.

Transparency is crucial and starts inside the government [32]. When the first outbreak began, the public reacted to transparency, comprehensive and timely information and data provided. Withholding the information created a vicious cycle of mistrust in authorities [33].

Clear direction in mitigating the outbreak is a must. The accurate and reliable information are underpinned to ensure the containment effort is aggressively used and disseminated at any form of social media to alert the public. In the absence of facts and trust, rumours and panic are inevitable as people turned out to be emotional. These emotions created anger and fear that posed a threat to the government. According to Amesh Adalja, a senior scholar at the Johns Hopkins Center for Health Security, argued that the public is discouraged to view the response from the government as a failure as that could create distrust of future public health measures [34].

At the geopolitical level, travel bans, border closing, trade controls, and others are other containment measures to urgently enhance global readiness needed in response to Covid-19. The consequences of outbreaks and epidemics are not distributed equally and the economic impact is also highly uncertain. Some sectors may even benefit financially, while others will suffer disproportionately [35]. Most countries that are badly affected by Covid-19 have prepared the stimulus package or domestic economic growth plan in order to boost spending activities.

According to Wei Yao, the chief Asia economist at Société Générale has stated that the worse the economic data are right now, the more aggressive policy responses will be. Given the current development, the infrastructure stimulus is almost certain and the question is just how much [36].

6. Conclusion

This research presented current trends of COVID-19 outbreak from 22nd Jan 2020 to 3rd March 2020 as visualized in CoronaTracker website. The trajectory of the outbreak is also forecasted by using SEIR model, 240 days from 20th January 2020. We also analyzed the sentiments from news extracted by CoronaTracker to further understand people's reaction towards this outbreak. COVID-19 is still an infectious disease with some unclear or unknown properties, which means accurate SEIR prediction can only be obtained once the outbreak has been successfully contained. The outbreak spreads are largely influenced by each country's policy and social responsibility.

In a pandemic like this, providing timely information to the public is paramount. A platform like CoronaTracker will assist the government and authorities to disseminate verified articles, provide updates to the situation, and advocate good personal hygiene to the people. CoronaTracker is built out of social responsibility to spread awareness to the common people by providing scientific-based data analysis, prediction and verified news. Our website has attracted more than 1000 users viewing our page at one time and surpassed 4 million-page views on 16th March 2020. Our platform is not limited to CoronaTracker webpage, but also Facebook page, Twitter and LinkedIn, with our ultimate goal is to provide verified news and current statistics to the world.

This paper is still an ongoing research as many more investigations regarding this disease can be carried out. Yet, it serves as the starting phase to research more in depth on questions that

revolve around this global pandemic.

Acknowledgments

This work is supported by Amazon AWS, LEAD and CirroLytix. We thank other CoronaTracker core members, Tan Wei Seng, Poon Chee Him and Marcus Chia for managing the community and the website. We also thank Yiran Jing, Zi Qing Ang, Goh Kok Han, Nikky Ng, Debby Huang, Shi Yong Pang, Roland Emmanuel Salunga and Kenny Kang for their involvement in our team discussion.

Author Group and Contributions

The CoronaTracker Research Team includes Fairoza Amira Binti Hamzah, Cher Han Lau, Hafeez Nazri, Dominic Ligot, Ummi Hasanah Zaidon, Mohammad Khursani Bin Mohd Shaib, Guanhua Lee, Tan Chen Liang, Adina Binti Abdullah, Chung Ming Hong, Chin Hwee Ong, and Chew Pei Ying. All team members jointly conceptualized the study, analyzed and interpreted that data, wrote and revised the manuscript and decided to submit for publication.

Corresponding author: Fairoza Amira Binti Hamzah, fairoza@kcg.ac.jp or fairozaamira@gmail.com .

References

[1] D. Hui et al, "The continuing 2019-nCoV epidemic threat of novel coronavirus to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China," *International Journal of Infectious Diseases,* vol. 91, pp. 264-266, 2020.

[2] World Health Organization (WHO), "Coronavirus disease 2019 (COVID-19) Situation

Report - 35," WHO, 2020.

[3] World Health Organization (WHO), "Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCov)," WHO, 2020.

[4] D. Toppenberg-Pejcic, J. Noyes, T. Allen, N. Alexander, M. Vanderford and G. Gamhewage, "Emergency Risk Communication: Lessons Learned from a Rapid Review of Recent Gray Literature on Ebola, Zika, and Yellow Fever," *Health Communication,* vol. 34, no. 4, pp. 437-455, 2018.

[5] L. Lin, R. McCloud, C. Bigman and K. Viswanath, "Tuning in and catching on? Examining the relationship between pandemic communication and awareness and knowledge of MERS in the USA," *Journal of Public Health,* p. fdw028, 2016.

[6] WHO, "Advice for Public," WHO Int., 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public. [Accessed 27 February 2020].

[7] World Health Organization, "Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)," World Health Organization, 2020.

[8] "Coronavirus Disease 2019 (COVID-19)," Centers for Disease Control and Prevention, 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/about/symptoms.html. [Accessed 27 February 2020].

[9] World Health Organization, "2019 Novel Coronavirus (2019-nCoV): Strategic Preparedness and Response Plan," World Health Organization, Geneva, 2020.

[10] World Health Organization, "Coronavirus disease 2019 (COVID-19) Situation Report- 13," World Health Organization, 2020.

[11] C. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *PLoS ONE,* vol. 5, no. 11, p. e14118, 2010.

[12] S. Oh, S. Lee and C. Han, "The Effects of Social Media Use on Preventive Behaviors during Infectious Disease Outbreaks: The Mediating Role of Self-relevant Emotions and Public Risk Perception," *Health Communication,* pp. 1-10, 2020.

[13] Kementerian Kesihatan Malaysia, "COVID-19," Kementerian Kesihatan Malaysia, 2020. [Online]. Available: https://www.facebook.com/kementeriankesihatanmalaysia/. [Accessed 27 February 2020].

[14] Noor Hisham Abdullah, "Noor Hisham Abdullah," Noor Hisham Abdullah, 2020. [Online]. Available: https://www.facebook.com/DGHisham/. [Accessed 27 February 2020].

[15] CoronaTracker Community , "CoronaTracker," CoronaTracker, 2020. [Online]. Available: https://www.coronatracker.com/. [Accessed 28 February 2020].

[16] A. Rachah and D. F. M. Torres, "Analysis, simulation and optimal control of a SEIR model for Ebola virus with demographic effects," *Commun. Fac. Sac. Univ. Ank. Series A1,* vol. 67, no. 1, pp. 179-197, 2018.

[17] A. T. Porter, "A path-specific approach to SEIR modeling," *Ph. D Thesis University of Iowa,* 2012.

[18] M. Munikar, S. Shakya and A. Shrestha, "Fine-grained Sentiment Classification using BERT," *ArXiv,* 2019.

[19] X. Dong and G. de Melo, "A Helping Hand: Transfer Learning for Deep Sentiment Analysis," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[20] U. Upadhyay, "Knowledge Distillation," Medium, 2018. [Online]. Available:

https://medium.com/neuralmachine/knowledge-distillation-dc241d7c2322. [Accessed 01 March 2020].

[21] John Hopkins University, "Coronavirus Map," John Hopkins University, 17 March 2020. [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed 17 March 2020].

[22] J. T. Wu, K. Leung, G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study," *The Lancet,* vol. 395, no. 10225, pp. 689-697, 2020.

[23] The Institute for Disease Modelling, "SEIR and SEIRs models," Institute for Disease Modelling, 2019. [Online]. Available: https://institutefordiseasemodeling.github.io/Documentation/general/model-seir.html. [Accessed 03 March 2020].

[24] Q. Li et al, "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia," *The New England Journal of Medicine,* pp. 1-9, 2020.

[25] R. Socher et al, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Conference on Empirical Methods in Natural Language Processing*, 2013.

[26] T. Wolf et al, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *ArXiv,* vol. abs/1910.03771, 2019.

[27] CoronaTracker Analytics Team, "Corona Tracker Analytics," CoronaTracker, 03 March 2020. [Online]. Available: https://www.coronatracker.com/analytics. [Accessed 03 March 2020].

[28] V. Agafonkin, "Leaflet," Leaflet, 2010. [Online]. Available: https://leafletjs.com/. [Accessed 7 March 2020].

[29] SciPy Developers, "SciPy," SciPy, 2020. [Online]. Available: https://www.scipy.org/.

[Accessed 13 03 2020].

[30] J. An, K. Ahn and M. Song, "Text Mining Driven Content Analysis of Ebola on News Media and Scientific Publications," *Journal of the Korean Society for Library and Information Science,* vol. 50, no. 2, pp. 289-307, 2020.

[31] Kim, Y. Jeong, Y. Kim, K. Kang and M. Song, "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news," *Journal of Information Science,* vol. 42, no. 6, pp. 763-781, 2016.

[32] S. K. Khor, "The Politics of the Coronavirus Outbreak," Think Global Health, 24 January 2020. [Online]. Available: https://www.thinkglobalhealth.org/article/politics-coronavirus-outbreak. [Accessed 04 March 2020].

[33] S. K. Khor, "Malaysia does not have a good record of transparency," The Star, 15 January 2020. [Online]. Available: https://www.thestar.com.my/opinion/columnists/vital-signs/2020/01/15/malaysia-does- not-have-a-good-record-of-transparency. [Accessed 04 March 2020].

[34] J. Passy, "Here's why the U. S. government's effort to contain the coronavirus outbreak was never going to be successful.," Market Watch, 29 February 2020. [Online]. Available: https://www.marketwatch.com/story/heres-why-the-coronavirus-may-spread-in-the-united-states-despite-government-efforts-to-contain-the-outbreak-2020-02-27. [Accessed 27 February 2020].

[35] D. E. Bloom, D. Cadarette and JP Sevilla, "New and resurgent infectious diseases can have far-reaching economic repercussions," *Finance and Development,* vol. 55, no. 2, pp. 46-49, 2018.

[36] M. McCormick, "Coronavirus: ECB ready to take "targeted" action to address economic

impact of outbreak - as it happened," Financial Times, 2 March 2020. [Online]. Available: https://www.ft.com/content/60f9e2ec-dd39-31cc-86d9-1adaa4dde1f8. [Accessed 04 March 2020].

APPENDIX: SEIR MODEL

In these equations, $S + E + I + R = N$ is the total population, with rate of spread, $\beta >$ 0, incubation rate $\sigma > 0$, and recovery rate $\gamma > 0$. The value of $\frac{dS}{dt}$ represents the rate of change S with respect to time t. Same as $\frac{dE}{dt}$, $\frac{dI}{dt}$, $\frac{dR}{dt}$.

The rate of spread, β is the rate of infection from an infected individual to one of their susceptible contacts on the unitary time step *dt*. For example, given two people A (infectious) and B (susceptible), the probability of B becoming infected after contacting A during the unitary time step is β. The term Δt is the difference between two observation points. Thus, number of individuals transferred from Susceptible state to Exposed state is

$$\frac{\beta SI}{N}\Delta t, \quad (14)$$

where is the $\frac{\beta SI}{N}$ is the Force of Infection in the SEIR model. Similarly, on the unitary time step, there are $\sigma E\Delta t$ number of cases transferred from Exposed state to Infectious, and $\gamma I\,(t)\Delta t$ number of cases transferred from Infectious to Removed.

Let *S(t), E(t), I(t)* and *R(t)* be the number of susceptible, exposed, infectious and removed individuals at time *t*, then

$$S(t + \Delta t) = S(t) - \frac{\beta S(t)I(t)}{N}\Delta t, \quad (15)$$

$$E(t + \Delta t) = E(t) + \frac{\beta S(t)I(t)\Delta t}{N} - \sigma E(t)\Delta t, \quad (16)$$

$$I(t + \Delta t) = I(t) + \sigma E(t)\Delta t - \gamma I(t)\Delta t, \quad (17)$$

$$R(t + \Delta t) = R(t) + \gamma I(t)\Delta t, \quad (18)$$

Based on the definition of the first-order derivative, $\frac{dX}{dt} = \frac{X(t+\Delta t)-X(t)}{\Delta t}$, as $\Delta t \to 0+$. Thus equation (15) – (18) can be rewritten as equation (10) – (13).

Assumptions of SEIR model [24] are as follows.

i. The SEIR model assumes a closed population, which means that the total number of populations is fixed, no births, no death, or introduction new individuals. From equation (10) – (13), we see that $\frac{d}{dt}[S(t) + E(t) + I(t) + R(t)] = 0$, where the population N is constant in any time t: $S(t) + E(t) + I(t) + R(t) = N$ for any $t \geq 0$.

ii. The individuals in the Exposed state are infected but not yet infectious.

iii. Well-mixed population.

iv. SEIR model assumes that the latent and infectious times of the pathogen are exponentially distributed.

In general, the dynamic SEIR is summarized as below [24].

a) Start the epidemic with a group of Susceptible individuals and at least one Infectious individual.

b) The Infectious individuals mix with the Susceptible class and create Exposed individuals following a probabilistic process.

c) Exposed individuals spend some days without spreading the virus and based on another probabilistic process become additional Infectious class.

d) Newly Infectious class repeat #2 and create more Exposed class.

Infectious individuals based on a probabilistic process either recover or die and become Removed class.