Determining the Effectiveness of the HEALTH-EDRM Framework and Twitter Data Analysis as the Near Real-Time Source of Information During the COVID-19 Situation

Mr. Kumpol Saengtabtim

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Industrial Engineering
Department of Industrial Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2020

การศึกษาประสิทธิผลของกรอบ HEALTH-EDRM และการวิเคราะห์ข้อมูลทวิตเตอร์ซึ่งเป็นแหล่งข้อมูลในเวลาใกล้เคียงความจริงในช่วงสถานการณ์โควิด-19

นายกำพล แสงทับทิม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมอุตสาหการ ภาควิชาวิศวกรรมอุตสาหการ
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

Thesis Title	Determining the Effectiveness of the HEALTH-EDRM Framework and Twitter Data Analysis as the Near Real-Time Source of Information During the COVID-19 Situation
By	Mr. Kumpol Saengtabtim
Field of Study	Industrial Engineering
Thesis Advisor	Assistant Professor Natt Leelawat, D.Eng.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

.... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

.... Chairman
(Associate Professor NARAGAIN PHUMCHUSRI, Ph.D.)

.... Thesis Advisor
(Assistant Professor Natt Leelawat, D.Eng.)

.... Examiner
(NANTACHAI KANTANANTHA, Ph.D.)

.... External Examiner
(Associate Professor Aussadavut Dumrongsiri, Ph.D.)

กำพล แสงทับทิม : การศึกษาประสิทธิผลของกรอบ HEALTH-EDRM และการวิเคราะห์ข้อมูลทวิตเตอร์ซึ่งเป็นแหล่งข้อมูลในเวลาใกล้เคียงความจริงในช่วงสถานการณ์โควิด-19. (Determining the Effectiveness of the HEALTH-EDRM Framework and Twitter Data Analysis as the Near Real-Time Source of Information During the COVID-19 Situation) อ.ที่ปรึกษาหลัก : ผศ. ดร.ณัฏฐ์ ลีละวัฒน์

การระบาดของวิกฤตโควิด-19 ถูกมองว่าเป็นหนึ่งในวิกฤตภัยที่มีความรุนแรงมากที่สุดซึ่งได้ทำให้เกิดผลกระทบทั้งชีวิตและเศรษฐกิจของประทั่วทั่วโลก กรอบบริหารความเสี่ยง HEALTH-EDRM นั้นถูกจัดทำขึ้นมาเป็นหลักการบริหารความเสี่ยงในช่วงปลายปี ค.ศ. 2019 จุดประสงค์หนึ่งของกรอบบริหารความเสี่ยงนี้คือการลดผลกระทบจากภัยต่าง ๆ ที่มีผลกระทบต่อชีวิตมนุษย์ การสื่อสารในช่วงวิกฤตนั้นก็ยังถือเป็นการทำงานตามหลักของกรอบบริหารความเสี่ยงนี้เนื่องจากสามารถช่วยลดผลกระทบของภัยที่เกิดขึ้นได้ ในงานศึกษานี้ การวิเคราะห์ประสิทธิผลของกรอบบริหารความเสี่ยง HEALTH-EDRM นั้นจะถูกนำมาวิเคราะห์โดยใช้ผลกระทบในเชิงจำนวนของผู้ติดเชื้อ, ผู้เสียชีวิต และอัตราส่วนของจำนวนผู้เสียชีวิตและติดเชื้อ จากวิกฤตโควิด-19 โดยใช้การวิเคราะห์ด้วยเครื่องมือ paired *t*-test เพื่อที่จะเปรียบเทียบประสิทธิผลของการปรับใช้กรอบบริหารความเสี่ยง HEALTH-EDRM ในกลุ่มประเทศสมาชิก และตามสภาวะเศรษฐกิจของกลุ่มประเทศที่อยู่ในบริเวณเอเชียโอเชียเนีย นอกเหนือจากนี้ทวิตเตอร์ซึ่งเป็นสื่อประเภทไมโครบล๊อกกิ้งนั้นก็จะถูกนำมาศึกษาเพื่อแสดงให้เห็นว่าสื่อประเภทนี้สามารถใช้สำหรับการสื่อสารในช่วงวิกฤตได้โดยการพิสูจน์จากคุณสมบัติความเร็วของข้อมูลตามกรอบเวลาที่ใกล้เคียงกับความจริงและความน่าเชื่อถือของข้อมูล จากในกรณีดังกล่าวการวิเคราะห์นี้จะใช้การเปรียบเทียบข้อมูลการรายงานสถานการณ์การระบาดของวิกฤตโควิด-19 ขององค์การอนามัยโลก และข้อมูลจากทวิตเตอร์ในช่วงวันที่ 21 มกราคม ค.ศ. 2020 ถึง วันที่ 16 สิงหาคม ค.ศ. 2020 ตามการวิเคราะห์ในกรอบเวลา รายวัน รายสัปดาห์ รายเดือน และรายไตรมาศ ซึ่งจะนำเอาหลักการหาความคล้ายของข้อมูลในเชิงตัวอักษรโดยใช้วิธี การจัดสรรของดีรีเคลแฝง หรือ Latent Dirichlet Allocation (LDA) และ ค่าความละม้ายโคไซน์ หรือ Cosine Similarity มาใช้ในการหาคำตอบ โดยผลสรุปของงานวิจัยนี้สามารถสรุปได้ว่าถึงแม้ว่ากรอบการบริหารความเสี่ยง Health-EDRM นั้นจะไม่สามารถนำมาช่วยกลุ่มประเทศสมาชิกได้มากในการจัดการวิกฤตโควิด-19 แต่กรอบบริหารความเสี่ยง Health-EDRM นั้นกลับมีประสิทธิผลอย่างมากในการช่วยจัดการวิกฤตโควิด-19 ในกลุ่มประเทศด้อยพัฒนา นอกจากนี้ ยังสามารถสรุปได้ว่าทวิตเตอร์นั้นมีความสามารถที่จะถูกนำมาใช้เป็นสื่อที่ใช้ในช่วงวิกฤตต่าง ๆ เนื่องจากความรวดเร็วตามกรอบเวลาใกล้เคียงกับความจริง และเป็นแหล่งข้อมูลที่มีความน่าเชื่อถือ

สาขาวิชา วิศวกรรมอุตสาหการ ลายมือชื่อนิสิต ....

ปีการศึกษา 2563 ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

6270015321 : MAJOR INDUSTRIAL ENGINEERING

KEYWORD: COVID-19 HEALTH-EDRM framework Content similarity Latent Dirichlet Allocation (LDA) Cosine similarity

Kumpol Saengtabtim : Determining the Effectiveness of the HEALTH-EDRM Framework and Twitter Data Analysis as the Near Real-Time Source of Information During the COVID-19 Situation. Advisor: Asst. Prof. Natt Leelawat, D.Eng.

Coronavirus pandemic or COVID-19 pandemic is considered to be one of the most severe disasters that affect both lives and economics to all the countries around the world. Health Emergency and Disaster Risk Management Framework or HEALTH-EDRM framework is the disaster management framework established in late 2019. One objective of this framework is to reduce the impact of the disaster that affects the lives of the people. Risk communication which is considered as the main function of this framework is also the main key for reducing the amount of affecting people. In this research, the analysis of the effectiveness of the HEALTH-EDRM framework will be defined based on the amount of the COVID-19 affected, death, and the ratio of death and affected cases in Asia Oceania. The paired *t*-test analysis will be used to compare the effectiveness for dealing with the COVID-19 situation in the aspect of the member of the framework and the economic situation. Furthermore, the Twitter microblogging platform will be defined as whether it can be used for risk communication function in terms of nearly real-time and reliable property. The analysis will be performed by comparing the content between the WHO’s situation report and the Twitter data from 21 January 2020 until 16 August 2020 based on daily, weekly, monthly, and quarterly based. The Latent Dirichlet Allocation (LDA) and Cosine Similarity will be used as the tools for generating topics and comparing the similarity. The results of both analyses show that even though the HEALTH-EDRM framework did not perform well to help their members for dealing with the COVID-19 situation, this framework can perform well for helping the least developed countries. In addition, Twitter is proved to be useful as a near real-time and reliable source of information.

Field of Study:	Industrial Engineering	Student's Signature ....
Academic Year:	2020	Advisor's Signature ....

ACKNOWLEDGEMENTS

For this research, I would like to thank Assistant Professor Natt Leelawat, who is my advisor. He is the one who provides advice about how to do the research. He is also the one who always supports me when I did my research in his laboratory.

In addition, this research cannot be finished without the comments and advice from Dr. Jing Tang. She is the one who taught me everything about the complex skill for performing the data analysis since I was in bachelor’s degree at SIIT, Thammasat.

Furthermore, I would like to thank all of the committee members, including Associate Professor Naragain Phumchusri, Dr. Nantachai Kantanantha, and Associate Professor Aussadavut Dumrongsiri based on their comments and advice based on my thesis.

Last but not least, this research cannot be finished without the support from my family, friends, and senior.

Kumpol Saengtabtim

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Glossary

No.	Word	Definition
1	Asia Oceania	Sixty-two countries that located in the area surrounded by India and Pacific oceans. (IFLA, 2017)
2	Content similarity	“To measure the degree of semantic equivalence between pairs of sentences”. (Bär, Biemann, Gurevych, and Zesch ,2012, p. 435)
3	Corpus	“A collection of documents on which retrieval is performed”. (Lin & Wilbur, 2007, p. 424)
4	Cosine similarity	“The inner product of two vectors divided by the product of their lengths”.(Ye, 2011, p. 91)
5	COVID-19	“The respiratory disease that caused by a novel coronavirus that is structurally related to the virus that causes severe acute respiratory syndrome”. (SARS)(Fauci, Lane, & Redfield, 2020, p. 1268)
6	Disaster	“Sudden unforeseen events with natural, technological or social causes that lead to destruction, loss and damage”. (Al-Dahash, Thayaparan, and Kulatunga ,2016, p. 1192)
7	HEALTH-EDRM Framework	“A risk-based approach to identify levels and causes of risk and how to mitigate them; a comprehensive scope incorporating prevention, preparedness, response, and recovery; including all hazards rather than fragmenting separate approaches for individual hazards; promoting multisectoral and multidisciplinary collaboration; having an inclusive people- and community-centered orientation; using the whole health system; and, being guided by ethical principles”.(Peters, Hanssen, Gutierrez, Abrahams, & Nyenswah, 2019, p. 317)
8	MERS	“An emerging coronavirus involved in severe acute respiratory distress syndrome (ARDS) and rapid renal failure”. (Poissy et al., 2014, p. 275)
9	Microblogging	“A new form of communication in which users can describe their current status in short posts distributed by instant messages, mobile phones, email or the Web”. (Java, Song, Finin, & Tseng, p. 56)

Continue

No.	Word	Definition
10	NLP	"A theory-motivated range of computational techniques for the automatic analysis and representation of human language". (Cambria & White, 2014, p. 48)
11	SARS	"A viral respiratory disease of zoonotic origin caused by the SARS-CoV". (Ryu, 2016, p. 181)
12	Semantic	"A large-scale study of conceptual structure and its lexical and syntactic expression in English". (Jackendoff, 1992).
13	Stemming	"A procedure to reduce all words with the same stem to a common form". (Lovins, 1968, p. 1)
14	Text mining	"The process of extracting interesting and non-trivial patterns or knowledge from text documents". (Tan ,1999, p. 65)
15	TF-IDF	"One of the most commonly used term weighting schemes in today's information retrieval systems". (Aizawa, 2003, p. 45)
16	Tokenizing	"The process of breaking a stream of text up into phrases, words, symbols, or other meaningful elements called tokens". (Verma, Renu, & Gaur, 2014, p. 16)
17	Tweet	"Message from the Twitter users, which can be posted up to 140 characters". (Boyd, Golder, and Lotan ,2010, p. 1)
18	Twitter	"A Social media giant famous for the exchange of short, 140-character messages called "tweet"".(Morstatter, Pfeffer, Liu, & Carley, 2013, p. 1)
19	Twitter API	"The mean to get at the rich Tweets and to build web application".(Makice, 2009, p. XII)

Chapter 1 Introduction

This chapter is the chapter for describing how this research is needed to be conducted based on the current COVID-19 situation. Based on this, this chapter includes five main parts that are the section 1.1 Background and significance of the problem, section 1.2 Research objectives, section 1.3 Scope of the study, section 1.4 Expected benefit, and section 1.5 Research process.

1.1 Background and significance of the problem

COVID-19, or Corona Virus Disease 2019, has significantly affected the world population since the start of the year 2020. On 30 June 2020, this disease had affected 10,185,374 people and killed 503,862 people worldwide; as shown in Figure 1 and Figure 2, COVID-19 was first known as pneumonia disease. Later on, this disease was called 2019-nCoV, named by The Coronaviridae Study Group (CSG) of the International Committee on Taxonomy of Viruses (of the International, 2020). This disease continued to attack the world population in many regions, such as Asia, Europe, and America. On 12 February 2020, WHO or World Health Organization announced the recent name for Coronavirus disease, which is COVID-19 or coronavirus disease 2019. After that, the ICTV or International Committee on Taxonomy of Virus explained that SARs-CoV2, or Severe Acute Respiratory Syndrome Coronavirus 2, was the virus that caused COVID-19 (L.-s. Wang, Wang, Ye, & Liu, 2020).

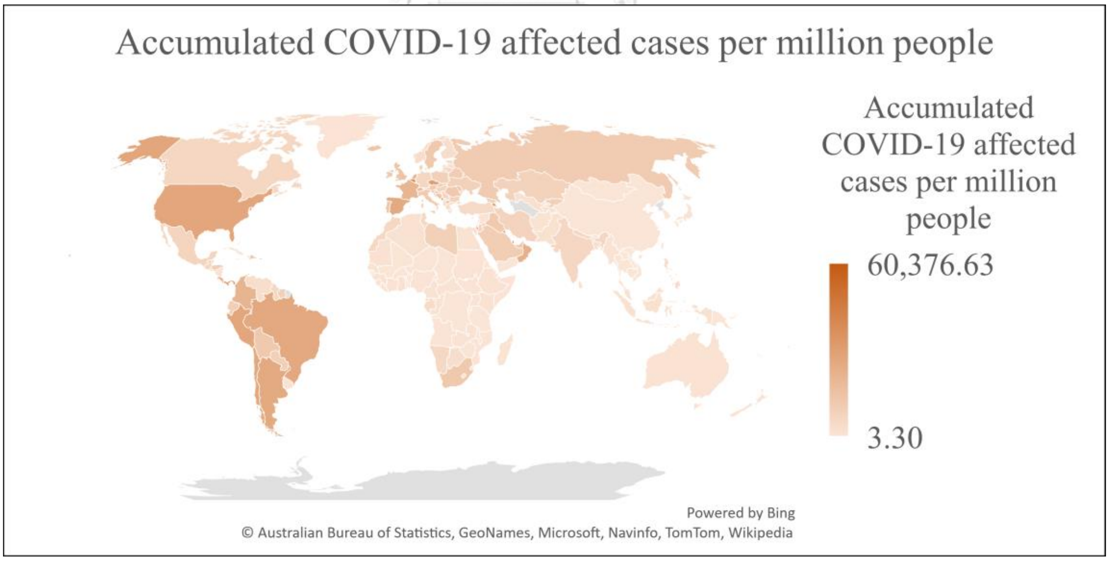

Accumulated COVID-19 affected cases per million people
Accumulated COVID-19 affected cases per million people
60,376.63
3.30
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Figure 1 Number of accumulated COVID-19 affected cases on 30 June 2020
Note. Source "Coronavirus Pandemic (COVID-19). Our World in Data," by Roser et al., (2020)

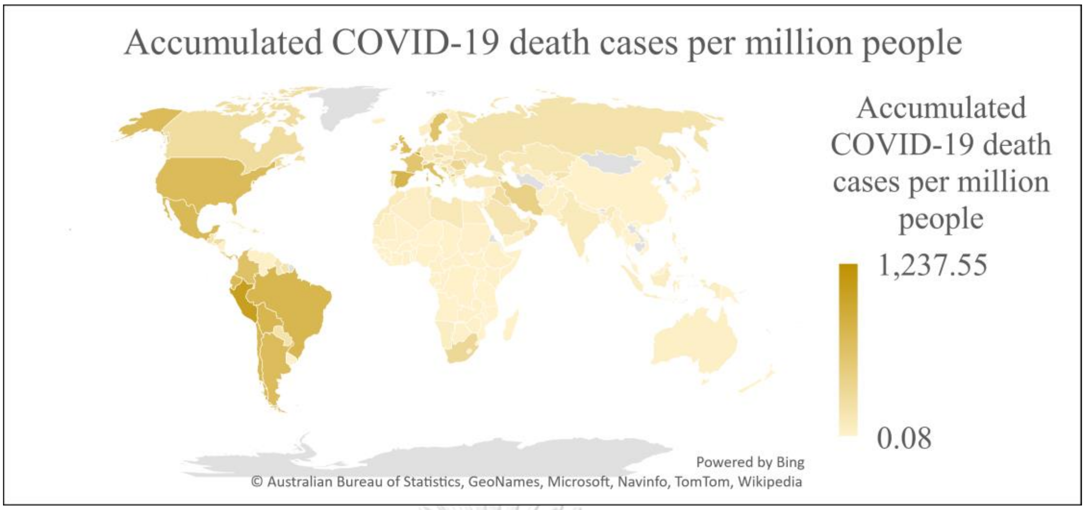

Accumulated COVID-19 death cases per million people
Accumulated COVID-19 death cases per million people
1,237.55
0.08
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Figure 2 Number of accumulated COVID-19 death cases on 30 June 2020
Note. Source “Coronavirus Pandemic (COVID-19). Our World in Data,” by Roser et al., (2020)

The number of confirmed cases regarding COVID-19 started to increase around the starting of March 2020 exponentially. Based on this, the statistical of the affected cases of the COVID-19 can be express. There were a lot of cancellations for events such as sports competitions, exhibitions, etc. Moreover, mass mask panic also occurred in the COVID-19 situation (Leung, Lam, & Cheng, 2020). This problem resulted from the shortage of supplies for surgical masks in many countries and territories such as Hongkong. The shortage problem did not affect only the supply for the surgical masks, but the supply of the alcohol gel or even the supply foods were also affected in the large areas. Many countries and territories tried to close their countries and territories and announce the policy for their people to self-quarantine. The result of this policy was the creation of new terms such as “Social distancing” or “Home isolation.” Many businesses and people also got a lot of economic effects based on COVID-19. Baldwin said that for the best-case scenario, the world GDP would reduce by 0.75% compared to the based line GDP. On the other hand, the world GDP will decrease by 1.75 % for the worst-case scenario (Baldwin & Weder di Mauro, 2020).

As the information about the COVID-19 that was previously identified, Table 1 shows the number of affected cases and death cases separated by the continent around the world. Based on this, the data retrieved from owid-19 and worldmeter (Roser, Ritchie, Ortiz-Ospina, & Hasell, 2020; Worldometers, 24 September, 2020). From Table 1, the most challenging situation is for the North American continent, and the least difficult situation is on the Oceania continent. In this research, the research will be focusing on the continents of Asia and Oceania. The reason why these two continents are the main focus is that the origin of the COVID-19 situation is in this continent, and some of the countries and territories in these continents, such as the Lao People's Democratic Republic, had reported 0 new affected cases for almost three months.

Table 1 Number of total affected cases and death cases for all continents

Continents	**Total affected cases (People)**	**Total death cases (People)**
Africa	393,444	9,878
Asia	2,259,222	55,858
Europe	2,357,080	187,538
North America	3,045,014	165,087
Oceania	9,344	133
South America	2,181,049	83,585

Note. Adopted from "Coronavirus Pandemic (COVID-19). Our World in Data," by Roser et al. (2020)

Since the end of the year 2019, there was a Health-related framework that was come out for reducing the impacts and losses from the multi disaster. This framework was called the Health Emergency and Disaster Risk Management Framework or HEALTH-EDRM framework (World Health, 2019). This framework was initially developed by the World Health Organization or WHO in October 2019. It is the framework that combines all the knowledge from all of the related people and organizations such as the members of WHO, the experts from the member states, and some of the partners who are involved in developing this framework. Based on this, the reason behind the development of this new framework is that even the country still has a disaster reduction framework such as International Health Regulations or IHR, etc., that country still has some weaknesses based on the multi disasters. Nowadays, the classification of the hazard has been divided into six main categories (Palliyaguru, Amaratunga, & Baldry, 2014), and it can be shown in Table 2. From Table 2, 6 types of hazards are identified. Based on these six types, the biological hazard is the main focus of this research because the example of the biological risks that have been identified are animal and human epidemics, which are related to the issue of the COVID-19 situation. In the past, several hazards similar to the COVID-19 had caused a lot of impacts to humanity, such as the Pandemic influenza in the year 1918, 1957, and 2009, Severe Acute Respiratory Syndrome or SARS in the year 2002, Middle East respiratory syndrome or MERS in the year of 2012, Ebola virus disease or EVD in the year of 2014, etc. Some of the frameworks related to risk reduction might be focused on some specific kinds of disasters. Based on this, the HEALTH-EDRM framework was developed to fulfill the objectives of the previous disaster and emergency mitigation framework by not only focusing on the specific kind of disaster but also focusing on multiple kinds of disasters and preparedness based on the disasters situations that might happen soon.

Table 2 Classification of hazard

No.	Types of hazards
1	“Geological hazards”
2	“Water and climatic hazards”
3	“Environmental hazards”
4	“Biological hazards”
5	“Chemical, industrial, and nuclear accidents”
6	“Accident-related hazards”

Note. Adopted from “Constructing a holistic approach to disaster risk reduction: the significance of focusing on vulnerability reduction,” by Palliyaguru, Amaratunga, & Baldry (2014), p. 48

Twitter is one of the most popular microblogging and social media platforms (Sakaki, Okazaki, and Matsuo .,2020). Most of the news and reports were generated by Twitter users all around the world. Some of the studies specify that Twitter acts as a real-time source of information(H. Wang, Can, Kazemzadeh, Bar, and Narayanan .,2012) (Sakaki et al. .,2020). In addition, Twitter can also be used as the trend generating (Mathioudakis & Koudas) to identify emergency cases such as disaster situations, news events, etc. During the period of the COVID-19 situation, there were a lot of tweets related to the COVID-19 situation that had been published to the public every day. Chen (Chen, Lerman, & Ferrara, 2020a) study about 123,113,914 tweets during the period of 21 January 2020 until 8 May 2020. Based on this, around 65.55% or 80,698,556 tweets of these tweets were in the English language. The research of Leelawat also mentions that the keywords related to the COVID-19 situation are nCoV, COVID-19, Coronavirus, etc. However, it depends on the nature of the language. For the Japanese language, it is specified that the keyword “コロナウイルス” is used as the main keyword for announcing the news and situation related to the COVID-19 situation in Japan(Leelawat, Tang, Saengtabtim, & Laosunthara, 2020).

In this research, around 60,000 English tweet data associated with the COVID-19 situation had been retrieved from 21 January 2020 until 30 September 2020. Twitter API was used as a tool for data collecting. The Tweepy library is the library for Twitter API that had been created by Roesslein (Roesslein, 2015). Initially, the researcher wants to see the characteristic of the data set by conducting a descriptive experiment by looking for the keywords that have been used during the COVID-19 situation. Based on this, the preliminary analysis shows some impressive results about the COVID-19 situation based on the frequency of the term that has been shown in Figures 3,4, and 5.

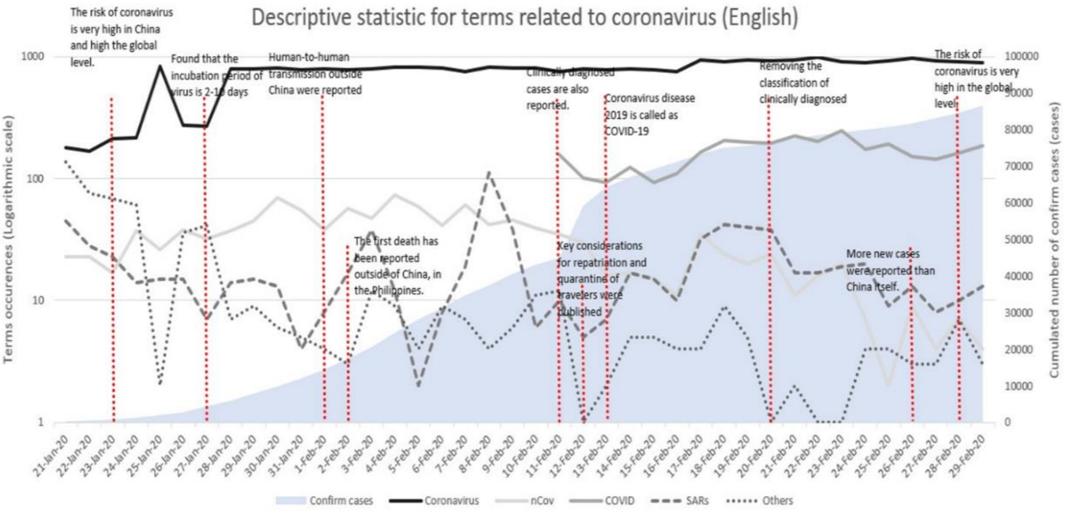


Figure 3 Descriptive statistic for terms related to coronavirus

Note. Adapted from “Novel Coronavirus (2019-nCoV): situation report, January to March,” by World Health Organization (2020)

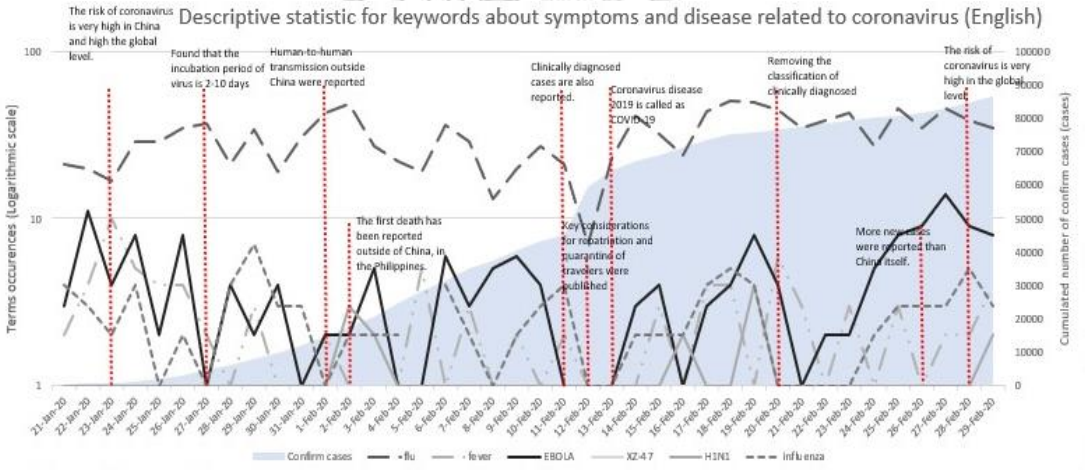


Figure 4 Descriptive statistic for keywords about symptoms and disease-related to coronavirus

Note. Adapted from “Novel Coronavirus (2019-nCoV): situation report, January to March,” by World Health Organization (2020)

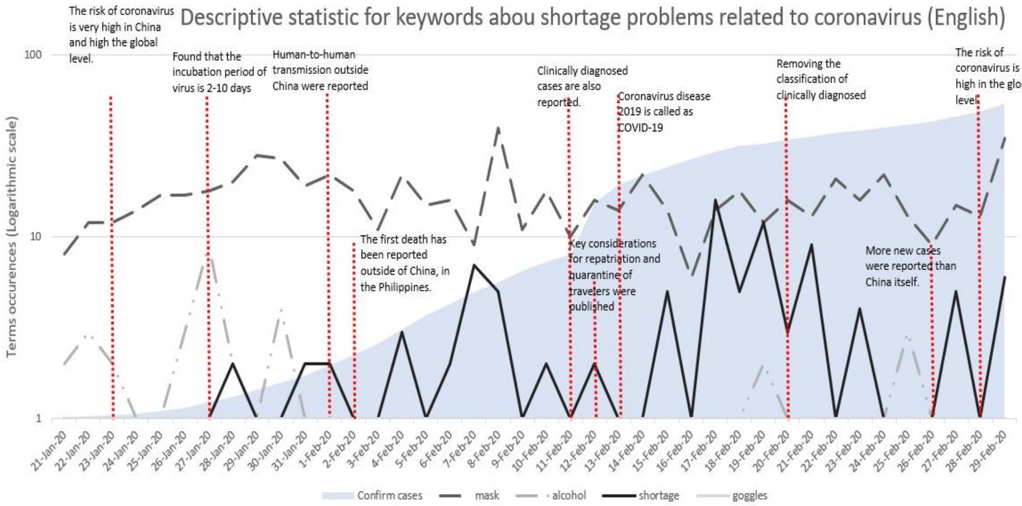


Figure 5 Descriptive statistic for keywords about shortage problems related to coronavirus

Note. Adapted from "Novel Coronavirus (2019-nCoV): situation report, January to March," by World Health Organization (2020)

Based on the preliminary analysis for the Twitter data, The content from the Twitter data also shows some relevant contents to the WHO situation reports, as shown in Figures 3,4 and 5. For example, the keywords of the COVID-19 also appear in both WHO's situation report and also Twitter data. In addition, some of the world COVID-19 effects problems such as the shortage problem for medical equipment also appear in both data set. However, there are also still have some flaws of the problem related to the difference announcing time between the WHO's situation reports and the Twitter data. Normally, the WHO's situation report will be announced at 10 AM (GMT+1) (World Health, 2020), and for Twitter, data will be ranged from the period of 12 A.M. until 12 P.M (GMT+7). Based on this, this analysis for finding whether the difference of the announcing time has some effects on the differences in the contents of both Twitter data and WHO's situation report or not.

1.2 Research objectives

The two main objectives will be solved based on this research topic. The first objective of this project is to find the effectiveness of the HEALTH-EDRM framework based on the COVID-19 situation in the area of Asia Oceania countries and territories, which can be illustrated by Figure 6. The effectiveness of the HEALTH-EDRM is how the member states or the countries and territories that applied the HEALTH-EDRM framework can act to reduce the impact of the COVID-19 situation. Based on this, it can be measured by using the number of accumulated total affected cases, death cases, and etc. For the second objective of this project, the determination of the nearly real-time manner of Twitter social media will be defined by using the tweets and comparing them with the reliable sources of news information.

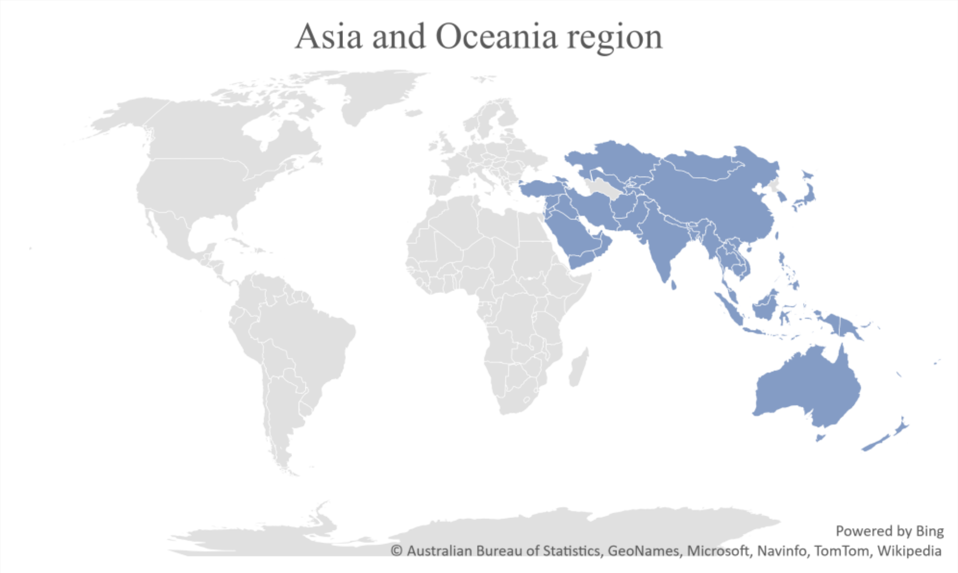

Asia and Oceania region
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Figure 6 Asia and Oceania region

1.3 Scope of the study

The scopes of this research are explained by the following statement.

1. The areas of 57 Asia Oceania countries and territories are the main focusing areas for the study countries and territories of the HEALTH-EDRM framework because this continent is the initial continent that COVID-19 impact the world before it was spreading to the other continents.
2. The COVID-19 tweets were started to collect from 21 January 2020 until 30 September 2020. Due to the limitation of the data storage and the performance of the computer processor, the Twitter data were collected by the amount of 1,000 tweets per day, and all of the tweets are in the English language. The initial day where the tweets are collected was the same day that the WHO has initially launched the first situation report for the COVID-19 situation, and the data were collected until 30 September 2020 because the researcher wants to investigate the situation for the COVID-19 for the semi-annual time frame since the COVID-19 had begun to outbreak.
3. This study aims only to focus on determining the effectiveness of the HEALTH-EDRM framework and the reliableness and real-time characteristics of the tweets during the COVID-19 situation.

The hypothesizes from the two objectives of this research are the study countries and territories of the HEALTH-EDRM framework will have a lower impact from COVID-19 than the countries or territories that did not be study countries and territories of the HEALTH-EDRM framework. For the second objective, the hypothesis of this objective is the Twitter social media can be proposed as a nearly-real time source of information during the disaster period.

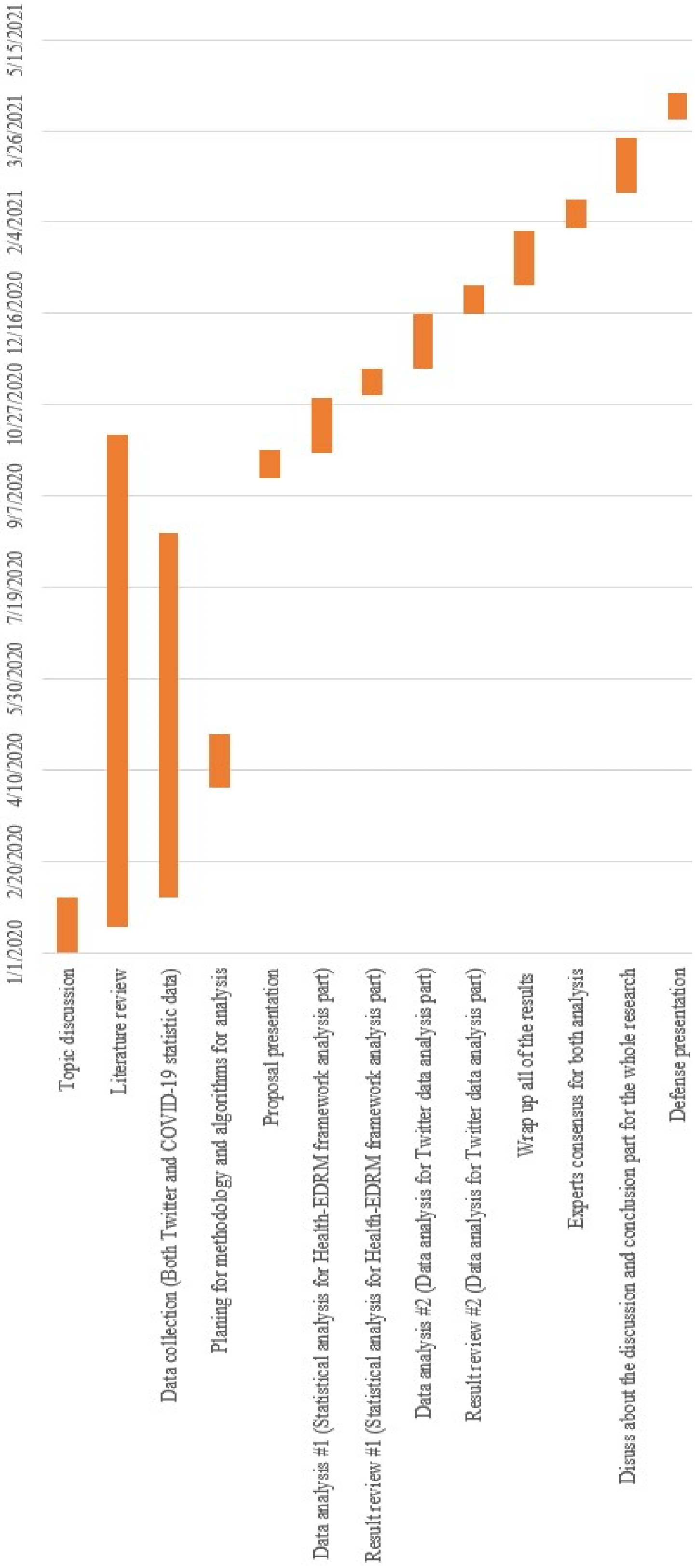
1.4 Expected benefits

1. To propose an effective health and disaster management framework to the countries or territories that has a high probability of disaster situations such as the Pandemic outbreak disaster.
2. To suggest the usage of the Twitter microblogging social media platform for both emergency alertness and information retrieval during the time of the COVID-19 situation.

1.5 Research process

For the research process in this study, most of the time was spent with the part of the literature review and data collection. Based on this, the timeline of this research can be explained in Table 3 as a Gantt chart. The research was initially conducted at the beginning of 2020. The main idea of this research came from the COVID-19 situation that has been impacted by people all around the world from January 2020 until now. In the initial phase, the researcher spent most of the time collecting data for performing the research analysis together with performing the literature review. After that, the research analysis will be performed after the proposal presentation. During the time of the research analysis, the analysis of this research will be performed for one objective at a time. Therefore, there will be 2 phases for the analysis part of this research. Finally, this research is planned to be finished by the beginning of April 2021.

Table 3 Gantt chart



Chapter 2 Literature review

In this chapter, the related concepts and the useful methods are concluded and briefly explained. The contents of this chapter are separated into five main parts. The first part is about the facts and news related to the COVID-19 situation that has impacted people all around the world (Section 2.1). The second part of this research is about the latest health and disaster-related framework that had been published by the WHO and the historical background of the disaster management framework (Section 2.2). The third part is about Information and communication from social media during disaster time (Section 2.3). The last three parts of this chapter are about Natural language processing and text mining (Section 2.4), Text Data preprocessing (Section 2.5), and Content similarity between the documents (Section 2.6).

2.1 COVID-19 situation and its impact around the world

The COVID-19 is the latest Pandemic outbreak that has impacted people all around the world. Previously, some diseases were closely related to the COVID-19 called SARS and MERS (Perlman, 2020), which was initially identified in 2003 and 2012, respectively. COVID-19 disease itself has been initially identified at the end of December 2019 in Wuhan city and began to spread from China and affected people all around the world (Guan et al., 2020). Initially, only a few studies are focusing on the data analysis for the COVID-19 situation due to the limitation of the data and a short study period. Most of the studies were focusing on the trend of the COVID-19 the root causes of the disease. For the research related to the trend of the COVID-19, the trend of the COVID-19 can be identified based on the summary of patient characteristics, an examination of age distributions and sex ratios, calculation of case fatality and mortality rates, geo-temporal analysis of viral spread, epidemiological curve construction (Novel, 2020). Similarly, Li also performs the analysis based on the COVID-19 statistical data, such as the report cases of the COVID-19, to find the trend of the COVID-19 in China (Q. Li et al., 2020). The mortality due to the COVID-19 is also focused on Verity (Verity et al., 2020). In this study, the range of the age can be identified based on the mortality rate due to the COVID-19 based on each range of ages. Apart from the researches related to the trends prediction and prediction of mortality rates based on the COVID-19 cases, the research related to the finding of the trend related to the news and information is also the topic that many researchers are focusing on Leelawat et al. defined that the trend of the COVID-19 situation can be defined based on the terms and keywords that have been specified in the information from Twitter (Leelawat et al., 2020). Not only the trend of the information can be used for understanding the situation for the COVID-19, but the trend of the information can also be used to find the feeling of the people during the COVID-19 situation too. Based on Zhang also shows that the trend of the information can also be used for understanding the emotion, which is mostly about depression, for the people during the time of COVID-19 situation (Y. Zhang et al., 2020). In addition, the mobility usage data was also used to find the trend for COVID-19 affected cases in China based on finding the correlation of the transportation usage and the number of affected cases in China (Kraemer et al., 2020). Apart from this, the *t*-test analysis, which is a popular statistical analysis method, has also be used in some researches related to the COVID-19 situation. Islam had applied the *t*-Test to analyze the factor that has some effect related to mental stress for the Bangladesh population (Islam, Bodrud-Doza, Khan, Haque, & Mamun, 2020). Similarly, the paired *t*-test has also be used for finding the difference in sentimental of the people for the time before and during the COVID-19 situation (S. Li, Wang, Xue, Zhao, & Zhu, 2020).

2.2 HEALTH-EDRM framework and Historical background of the Disaster management framework

Currently, many organizations and countries are trying to concern about health and disaster-related problems. There are also a lot of disasters and risk-related frameworks that had been created by world organizations such as the United nation (UN), the World Health Organization (WHO), etc. For the past 20 years from 2020, the disasters and risk-related frameworks are shown in Figure 7.

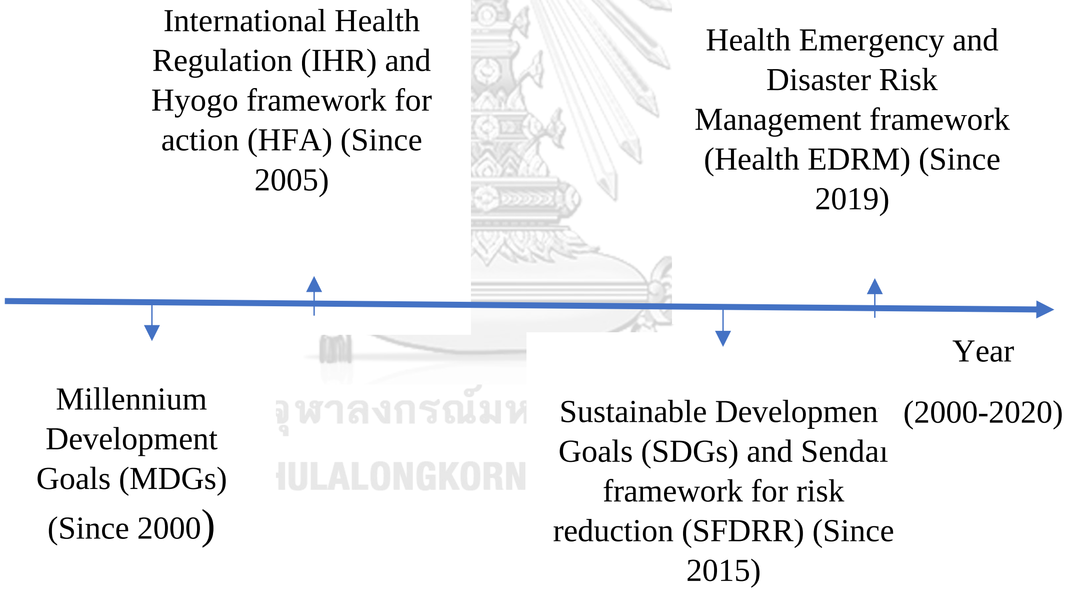

International Health Regulation (IHR) and Hyogo framework for action (HFA) (Since 2005)
Health Emergency and Disaster Risk Management framework (Health EDRM) (Since 2019)
Year
(2000-2020)
Millennium Development Goals (MDGs) (Since 2000)
Sustainable Developmen Goals (SDGs) and Sendaı framework for risk reduction (SFDRR) (Since 2015)

Figure 7 Timeline for the disasters and risk-related framework

For the initial stage of development goal since September 2000, the Millenium Development Goals or MDGs had been proposed by the UN. For the proposed goals, eight actions were needed to be done (World Health, 2015) for developing the member countries. These eight actions are expressed in Table 4.

Table 4 Action for fulfilling the objective for MDGs

No.	**Action**
1	“Eradicating extreme poverty and hunger”
2	“Achieving universal primary education”
3	“Promoting gender equality and empower women”
4	“Reducing child mortality”
5	“Improving maternal health”
6	“Combating HIV/AIDS, malaria, and other diseases”
7	“Ensuring environmental sustainability”
8	“Developing a global partnership for development”

Note Adopted from “Health in 2015: from MDGs, millennium development goals to SDGs, sustainable development goals,” by World Health Organization (2015), p.4.

Based on the needed action for MDGs, all of them are important for achieving the highest goals for developing countries. Among these eight actions, 3 of them are related to health issues. These three health issues are reducing child mortality health, improving maternal health, and combating HIV/AIDS, malaria, and other diseases (World Health, 2015). After 15 years since 2000, Sustainable development goals or SDGs had been proposed by the UN. Based on the concept of the SDGs, SDGs have the objectives to create sustainable growth based on the three aspects are economic, social, and environmental. Among these three aspects, the Health-related issue is also the main focus of the SDGs (World Health, 2016). SDGs came up with 17 sustainable development goals to cope with the three aspects that have been previously defined. These 17 goals can be expressed in Table 5.

Table 5 Sustainable development goals

No.	**Goals**	**No.**	**Goals**
1	“No Poverty”	10	“Reduced Inequality”
2	“Zero Hunger”	11	“Sustainable Cities and Communities”
3	“Good Health and Well-being”	12	“Responsible Consumption and Production”
4	“Quality Education”	13	“Climate Action”
5	“Gender Equality”	14	“Life Below Water”
6	“Clean Water and Sanitation”	15	“Life on Land”
7	“Affordable and Clean Energy”	16	“Peace and Justice Strong Institutions”
8	“Decent Work and Economic Growth”	17	“Partnerships to Achieve the Goal”
9	“Industry, Innovation, and Infrastructure”		

Note Adopted from “The Sustainable Development Goals and Addressing Statelessness,” by UN High Commissioner for Refugees (UNHCR) (March 2017) p. 1.

According to the 17 goals of the SDGs, Murray identified that there are seven goals related to the health of the people (Murray, 2015). The health-related goals are zero hunger, good health and well-being, Gender equality, Clean water and sanitation, Sustainable cities and communities, Climate Action, and Peace and strong institution. In contrast, Lim summarizes the health-related SDG goals into ten goals created by Inter-Agency and Expert Group (Lim et al., 2016). These ten goals are No poverty, Zero hunger, Good health and well-being, Gender Equality, Clean Water and Sanitation, Affordable and Clean Energy, Decent Work and Economic Growth, Sustainable Cities and Communities, Climate Action, and Peace and Justice Strong Institutions. According to these two studies, it can be defined that most of the goals of the SDG are related to the health condition of the people, which also included health-related diseases like Pandemic outbreak too.

According to the third target of Sustainable Development Goals (SDG), the health of the people is the main focus of the United nation (*The Sustainable Development Goals and Addressing Statelessness*, March 2017). The latest disaster management framework, like Sendai Framework for Disaster Risk Reduction (SFDRR), also set the seven global targets (Unisdr ,2015), which also included the reduction in mortality rate and affected people based on the disaster. This framework was created four years after the 2011 Great East Japan tsunami. The targets and goals from these two examples, together with the Universal Health Coverage (UHC), International Health Regulation (IHR), and the Paris Agreement on Climate Change are the main concepts to build up the Health Emergency and Disaster Risk Management Framework (HEALTH-EDRM framework) (World Health, 2019). This framework was developed by WHO in October 2019. The objective of this framework is to prevent and protect the world population from disaster and health-related problems by creating a preparedness and readiness based on the concept of disaster management to the country that applied this framework. This framework points out ten functions and components for fulfilled its visions and objectives. The functions and components of this framework can be identified in Table 6. Based on the functions and components of the HEALTH-EDRM framework, it is required that all of the stakeholders of the health and disaster-related organization should take their responsibilities to obtain the highest effectiveness from this framework. Only a few pieces of research were supporting the benefits based on the HEALTH-EDRM framework since the HEALTH-EDRM framework is the framework that has just been initially launched since the end of 2019. There were also several pieces of research, such as the research of Djalante, Ishiwatari proposed that the HEALTH-EDRM can be utilized the practice and create a response to the COVID-19 situation (Djalante, Shaw, & DeWit, 2020) (Ishiwatari, Koike, Hiroki, Toda, & Katsube, 2020).

Table 6 Functions and components that the organization should have based on the HEALTH-EDRM framework

No.	**Functions and Components**	**Definition**
1	"Policies strategies and legislations."	The good use of the HEALTH-EDRM framework needs to support the national legislation and policies for each country
2	Planning and coordination	The HEALTH-EDRM will work and coordinate with the national IHR and Sendai framework
3	Human resource	All level of the related organization, including national, sub-national, and local level, should be involved to perform the key activities
4	Financial resource	The HEALTH-EDRM framework needs an adequate amount of budget from the government
5	Information and knowledge management	Information and knowledge management is the main key to reduce the impact based on the disaster
6	Risk communication	The critical function of the HEALTH-EDRM framework, which also needs real-time access for the information
7	Health infrastructure and logistic	Well prepared for health facilities is the key to reducing the losses from the disaster
8	Health and related services	Public health and clinic should also prepare for emergencies cases
9	Community capacity for HEALTH-EDRM	The help of the community can help the HEALTH-EDRM framework to identify the risk and vulnerability
10	Monitoring and evaluation	A Health monitoring system should be used to identify risk and understand the situation more clearly

Note. Adapted from "Health emergency and disaster risk management framework," by World Health Organization (2019) p. 9-11.

2.3 Information and communication from social media during disaster time

At present, Social media have become a powerful tool for various purposes (Kongthon, Haruechaiyasak, Pailai, & Kongyoung, 2014). Similarly, Scott and Errett also defined that social media are currently used as the main tool for disaster response and communication (Scott & Errett, 2018). Kongthon analyzes the uses of social media called Twitter during the time of the 2011 Thai flood for the purpose of spreading the emergency and information. Also, Scott and Errett analyze the uses of social media such as Facebook and Twitter during the time of the 2016 Louisiana floods (Scott & Errett, 2018). Quarantelli defined that good or successful disaster management can be resulted based on the good activities of the emergency organization, which also included communication and information related (Quarantelli, 1988). In this study, the problem of communication can be identified based on the view of between organizations, from organizations to the public, from the public to organizations, and within systems of organizations (Quarantelli, 1988). Social media also be used to analyze public behavior during the disaster period (Chae et al., 2014). In Chae, it is also defined that public behavior is also related to the disaster response, which is the key function of disaster management. Based on this, from the 6th function and component of the HEALTH-EDRM framework (World Health, 2019) that has been previously identified, it defined that effective communication and risk communication is the key function for the health emergency and disaster risk management. In addition, real-time information is required to be obtained by everyone to make the correct decision and reduce the loss based on the disaster.

2.4 Natural language processing and text mining

Text mining analysis is one of the tools for performing data analysis. In this case, text data is the main source of data for performing the analysis. Text data is defined as unstructured data (Buneman, Davidson, and Suciu , 1995). Based on this, the unstructured data is the data that does not have the model in itself, and this type of data can’t be put in the form of the table (Rusu et al. ,2013). At present, there are a lot of demands for the people who want to extract the information based on this type of unstructured data to be used in many organizations (Rusu et al. ,2013). In addition, due to the huge amount of text data in the current world, text data is the main focus for performing the analysis. Regarding the challenge of text extraction, the term text mining or text data mining (Tan ,1999) is represented as the main tool for solving this challenge. Text mining, in this case, is the process of extracting interest and knowledge based on unstructured data (Tan ,1999). The concept of text mining is the concept of focusing the text of the words as the string format, or it can be called the bag of words (Aggarwal & Zhai, 2012). Based on this, the real information of the words will not much be considered based on the process of text mining. The text information or text semantic is the developed view of the text data because, in this case, the meaning and the pattern of the text apart from just the bag of words are also included in the analysis (Aggarwal & Zhai, 2012). On the other hand, natural language processing or NLP is a closely related term to text mining, but it focuses more on the full meaning representation of the text (Kao & Poteet, 2007).

2.5 Text Data preprocessing

Before performing any data analysis, the step of data preprocessing should be the initial step after retrieving enough data and information. Based on Kannan and Gurusamy, this study suggests that the step of data preprocessing is the critical step (Kannan & Gurusamy, 2014). Similarly, Kalra and Aggarwal also suggest that data preprocessing is the subset of data preparation (Kalra and Aggarwal ,2017). The objective of the data preprocessing step is to reduce the amount of data and improve the efficiency of information retrieval (Kannan & Gurusamy, 2014), which is a term that indicates the method of retrieving the relevant data for analysis (Cooper, 1971). Similarly, Vijayarani states the definition of Information retrieval by means of retrieving text data from a large amount of text information (Vijayarani, Ilamathi, & Nithya, 2015). Based on this, the method for data preprocessing is differently explained by the previous research. Kalra and Aggarwal express the method for data preprocessing by using five main steps, which are Tokenizing, Filter stop word, Filter token by length, Stemming, and lastly, Transform cases. For tokenizing, it is the step that the terms of the sequence in the whole document are split into words. Based on this step, the punctuation in the documents will be removed.

Next, Filter Stop-Words; in this step, the unimportant terms such as WH-Question or helping verb of the document will be removed. For the Filter token by length step, in this step, some terms will be removed based on the specific range of the words. For the stemming, it is the step for changing some terms to the root-based form for each term. For the Transform cases step, this step is to convert all the terms to be either lower cases or upper cases. In contrast, Kannan and Gurusamy proposed only the three steps for the process of data preprocessing, which are Tokenization, Stop word removal, and stemming. All of the three methods in this study are most relevant to the method from the first study, but some of the steps in the second study are merged into other steps. At present, there are a lot of libraries that provide a useful function for performing data preprocessing. The Natural Language Toolkit or NLTK is the python library that is used for working with human language data (Bird, Klein, & Loper, 2008). This library tool was invented in 2002 by Looper and Bird (Loper & Bird, 2002). This library provides a variety of useful functions related to text mining, which include the function of data preprocessing. Another library that can perform the data preprocessing task is Scikit learn (Buitinck et al., 2013). This library not only provides the code for working with text mining but also provides tutorials and examples for each function inside this library too.

2.5 Topic generated algorithm

Nowadays, unstructured data appear to be important for many industries such as news agencies, the services business, etc. For the services business, the contents that the customers express for their satisfaction are the content that the services provider needs to extract from the huge amount of feedback and comments from the customers

(Kasper and Vela 2011). In addition, the news agencies or the organizations related to the news and information also still need to extract the news and useful information from the huge amount of the data of news too (Elliott, 1998). Currently, there is a lot of topic modeling algorithm that can generate the topic based on the huge amount of text such as Probabilistic Latent Semantic Analysis (PLSA), Biterm Topic Model (BTM), and Latent Dirichlet Allocation (LDA) (W. Li, Feng, Li, & Yu, 2016). Based on this, the LDA algorithm is currently the famous algorithm that can generate the topic from the high dimensional data, which will give useful and important information (Ramya, Sejal, Venugopal, Iyengar, and Patnaik ., 2018). In addition, it can also be used to extract the information based on the high dimensional data for reducing the complexity and redundancy (Ramya et al. ., 2018). From this, the LDA algorithm has been used for generating and extracting the important topic in many fields of application, such as clustering the topic based on the legal judgments (Raghuveer, 2012), generating the topic from the Microsoft Research Paraphrase Corpus (Rus, Niraula, and Banjade ., 2013), and etc.

2.6 Content similarity between the documents

In the past, there was a lot of researches that try to perform content analysis based on some sources of information during the disaster periods. The reason for this research is to understand the disaster situation clearer. Based on Meng, the information from the newspapers was used to perform the analysis for crisis management during the SARS epidemic in China(Meng & Berger, 2008). Similarly, the newspaper is also used as the source for performing the analysis based on the political aspect based on the event of the Haitian earthquake (Gurman & Ellenberger, 2015). Not only the earthquake or Pandemic disaster but for the tsunami and hurricane event, there is also the research that used the newspaper for performing the content analysis based on the death case due to Hurricane Katrina and the 2011 Great East Japan Tsunami. The content from the magazine was also used as the main source of information to perform the content analysis based on the trend of breast cancer for women (Lantz & Booth, 1998). The content from the report from the government was also be used to perform some content analysis too. Chatfield and Brajawidagda show that the content analysis based on the government report can be used to clarify the process for risk management during the tsunami disaster in Indonesia (Chatfield and Brajawidagda ,2013). This study also suggested that Twitter can also be used during a disaster period like a tsunami too. In addition, the uses of social media data and online media were also one of the main focuses of this research. Online forums are one of the online contents that have been used to perform the contents analysis. Tirkkonen and Luoma-aho uses the content from the two popular online fora in Finland to analyze the management of the related authorities based on the case of Swine flu (Tirkkonen & Luoma-aho, 2011). Apart from the content analysis, there was also a lot of research about finding the similarity of the content between the terms and contents of the two documents. A good point of content similarity Based on the research of Gomaa and Fahmy states that the text similarities can be divided into three main categories, which are string-based similarity, corpus-based similarity, and knowledge-based similarity

(Gomaa & Fahmy, 2013), which can be illustrated by Table 7. For the string-based similarity, this kind of measurement is the method for finding the similarity of the text based on the sequence of the terms and the characters. Based on this, the distance between the two terms will be calculated by transforming the two terms into the vector format (Vijaymeena & Kavitha, 2016). For the corpus-based similarity, it is the measurement of the similarity of the meaning or semantic of the contents by using the knowledge based on the big storage of data and information or corpus (MacMullen, 2003). For the knowledge-based similarity, it is the measurement of similarity based on the similarity between both of the contents and how the texts are arranged inside each document. Based on this, the comparison between the contents will be compared based on the semantic network or the network that identifies the words pattern and how they are connected (Simpson & Usey Jr, 2004).

Table 7 Text similarity algorithm techniques

Technique	Methodology
String-based similarity	"Block Distance, Cosine similarity, Dice's coefficient, Euclidean distance, Jaccard similarity, Matching Coefficient, Overlap coefficient"
Corpus-based similarity	"Hyperspace Analogue to Language (HAL), Latent Semantic Analysis (LSA) , Generalized Latent Semantic Analysis (GLSA), Explicit Semantic Analysis (ESA), The cross-language explicit semantic analysis (CLESA), Pointwise Mutual Information - Information Retrieval (PMI-IR), Second-order co-occurrence pointwise mutual information (SCO-PMI), Normalized Google Distance (NGD), Extracting DIStributionally similar words using COoccurrences (DISCO)"
Knowledge-based similarity	"Similarity and relatedness"

Note. Adopted from "A survey of text similarity approaches," by Gomaa, W. H., & Fahmy, A. A. (2013). International Journal of Computer Applications, 68(13), 13-18.

The Cosine similarity analysis, together with the TF-IDF technique, is going to be applied for finding the content similarity between the documents. The documents that are selected to be compared are the data from Twitter that is related to the COVID-19 situation and the data from the WHO's situation reports. The cosine measured originally popular in the field of mathematic. The cosine is the measurement method for finding the similarity between the two vectors (B. Li and Han ,2013). Similarly, text can be converted into a vector format. Based on this property, there are a lot of text analysis researches that used the Cosine similarity technique as a tool for measure the similarity between the contents between the documents. Some of the studies defined cosine similarity as a popular and robust tool for performing text analysis (B. Li and Han ,2013) (Tata & Patel, 2007). Furthermore, Cosine similarity has been used in many fields related to the content and similarity analysis, such as pattern recognition and medical analysis (Q. Li et al., 2020; Ye, 2011), measure the similarity between the three data set (Tata & Patel, 2007),text-independence speaker verification (Novoselov, Shchemelinin, Shulipa, Kozlov, and Kremnev ,2018), clustering content for a specific problem (Vijaymeena & Kavitha, 2016), etc.

Chapter 3 Research design and methodology

In this part, the research design and methodology of this research are going to be clarified. The research design mostly came from the literature review section and the inspiration based on the problem of the current COVID-19 situation. The first section of this chapter is the research design (Section 3.1). This section states how the research is going to be performed based on the two objectives. In section 3.2, research hypotheses, this section is mainly about the preliminary outcomes of the two objectives. Lastly methodology, in this section, the methodologies for solving the research objectives are explained.

3.1 Research design

Based on the literature review, the usefulness and benefit have been described based on the goal and vision of the HEALTH-EDRM framework. However, the effectiveness of this framework is still unclear and needs some analysis to see the outcome based on this framework. Based on this, determining the effectiveness of the HEALTH-EDRM framework during the COVID-19 situation can be done by using the index for measuring the severity based on the COVID-19 situation. The indexes that have been selected for analysis based on the literature review and the availability of the data are total accumulated COVID-19 affected cases per million people (APM), total accumulated COVID-19 death cases per million people (DPM), the ratio of COVID-19 death cases and accumulated affected cases (RDA).

3.1.1 Total accumulated COVID-19 affected cases per million people (APM)

The APM is the value that indicates the amount of the total accumulated COVID-19 cases since there is a case in that specific over the total population of that country and territory. Based on this, the more value of APM, the higher severity to that specific country or territory. The equation for calculating this index is defined by equation (3.1)

$$\text{Total accumulated COVID} - 19 \text{ affected cases per million people} \\ = \frac{\text{Total accumulated affected cases}}{\text{Total population}} * 1{,}000{,}000 \qquad (3.1)$$

3.1.2 Total accumulated COVID-19 death cases per million people (DPM)

The DPM is the value that indicates the amount of the accumulated COVID-19 death cases since there is a first death case in that specific over the total population of that country or territory. The equation to calculate this index is defined by equation (3.2)

$$\text{Total accumulated COVID} - 19 \text{ death cases per million people} \\ = \frac{\text{Total accumulated death cases}}{\text{Total population}} * 1{,}000{,}000 \qquad (3.2$$

3.1.3 The ratio of COVID-19 death cases and accumulated affected cases (RDA)

The RDA is the ratio for identifying the number of death cases due to the COVID-19 over the total accumulated affected cases of the COVID-19. Based on this, the proportion can indicate how worse the situation of the specific country or territory. This proportion can be expressed as equation (3.3)

$$\text{The ratio of COVID} - 19\text{ death and accumulated affected cases} \\ = \frac{\text{Total accumulated death cases}}{\text{Total accumulated affected cases}} \qquad (3.3)$$

Based on the three indexes that have been previously identified, these indexes can be used for analyzing the impact of the COVID-19 in the specific country and territory. The area of Asia Oceania has been selected because this area is the initial spreading source for the COVID-19 situation. In addition, around half of the members of the HEALTH-EDRM framework are also located in Asia and the Asia Oceania region. The members of the HEALTH-EDRM framework can be summarized in Table 8. After scoping down to the area of Asia Oceania region, the researcher also shows all the countries and territories in Asia Oceania that has been selected to analyze and those countries and territories that did not include due to the availability of the data. All of the information related to the study area can be summarized in Table 9.

Table 8 Members of HEALTH-EDRM and the Member state of HEALTH-EDRM in Asia Oceania

Member of HEALTH-EDRM framework	Member of HEALTH-EDRM framework in Asia Oceania
Australia, Bangladesh, Cambodia, Canada, China, Egypt, Ethiopia, India, Indonesia, Islamic Republic of Iran, Japan, Lao People's Democratic Republic, Mexico, New Zealand, Oman, Peru, Philippines, Qatar, Republic of Moldova, Singapore, Sri Lanka, Sudan, Turkey, United Kingdom, United Republic of Tanzania, United States of America (USA) and Viet Nam	Australia, Bangladesh, Cambodia, China, India, Indonesia, Iran, Japan, Laos, Oman, New Zealand, Qatar, Philippines, Singapore, Sri Lanka, Turkey, Vietnam

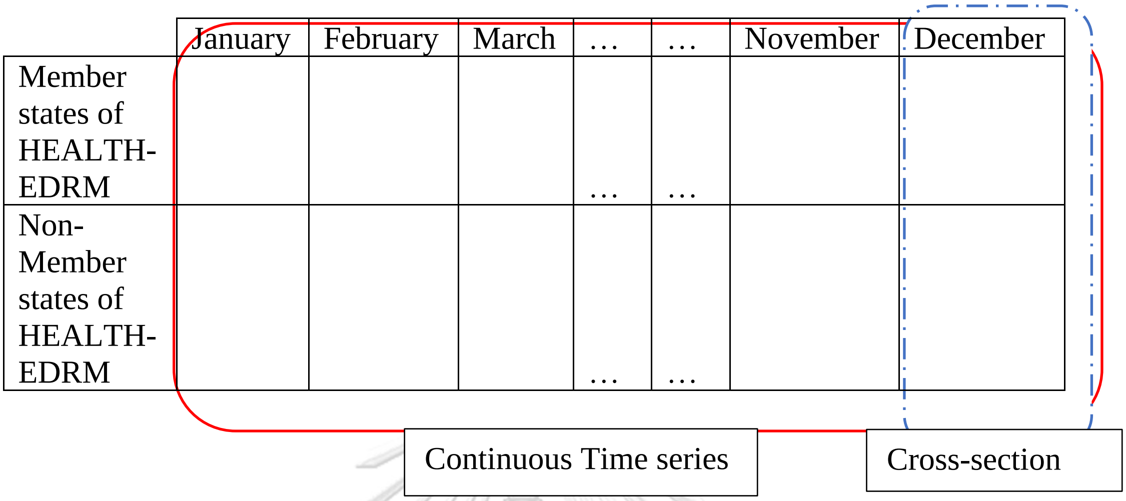
Note. Adopted from "Health emergency and disaster risk management framework," by World Health Organization (2019), p. vii.

Table 9 Countries and territories in Asia Oceania that has and has not been included in this research

Included countries and territories		Not included countries and territories
Afghanistan	New Zealand	North Korea
Armenia	Northern Mariana	Timor-Leste
Australia	Islands	Solomon Islands
Azerbaijan	Oman	Vanuatu
Bahrain	Pakistan	Samoa
Bangladesh	Palestine	Kiribati
Bhutan	Papua New Guinea	Micronesia
Brunei	Philippines	Tonga
Cambodia	Qatar	Marshall Islands
China	Saudi	American Samoa
Fiji	Arabia	Palau
French	Singapore	Cook Islands
Polynesia	South	Turkmenistan
Georgia	Korea	Tuvalu
Guam	Sri Lanka	Wallis and Futuna
Hong Kong	Syria	Nauru
India	Taiwan	Niue
Indonesia	Tajikistan	Tokelau
Iran	Thailand	
Iraq	Timor	
Israel	Turkey	
Japan	United	
Jordan	Arab	
Kazakhstan	Emirates	
Kuwait		
Kyrgyzstan		
Laos		
Lebanon		

Note. Adapted from "Health emergency and disaster risk management framework," by World Health Organization (2019), p. vii.

The analysis to identify the effectiveness of the HEALTH-EDRM will be performed based on the focusing area that has been previously identified. Based on the three indexes of the COVID-19 statistical data, the analysis for finding the difference between the value of the indexes between the member state of HEALTH-EDRM country and the country that aren't the member states of the HEALTH-EDRM will be performed. From the literature review, the analysis based on the continuous-time series of the COVID-19 data (Das, 1994) and the analysis based on the literature review related to COVID-19 cases in each country is a popular analysis method. Therefore, this research will perform the analysis to find the difference between the two groups of countries and territories in terms of the three indexes by using the times series and cross-section analysis. Based on this analysis, the method for finding the difference in terms of the mean will be performed. The use of *t*-Test analysis will be selected to perform the analysis to see the difference between the impact of the COVID-19 situation based on the time series and cross-section analysis, which can be illustrated by Figure 8.



	January	February	March	…	…	November	December
Member states of HEALTH-EDRM				…	…		
Non-Member states of HEALTH-EDRM				…	…		

Figure 8 Cross-section and continuous time series analysis

3.1.4 Cross-section analysis

For the cross-section analysis, the analysis will be performed based on a specific time since the COVID-19 has been started to impact the world population since December 2019 and tend to continue to impact the world population until the end of the year 2020. The cross-section analysis will be performed at the time of 30 2020. Based on this, the three focusing indexes to identify the impact of COVID-19 will be retrieved only on 30 September 2020 for performing the analysis. Therefore, the difference between the COVID-19 situation of each group of the country and territory can be seen based on a specific period.

3.1.5 Continuous Time series analysis

Apart from the cross-section analysis, the continuous-time series analysis is another method for analyzing the different situations based on a wide range of periods (Das, 1994). The analysis for finding the difference between the two groups of the countries and territories will be performed by comparing the mean of the three indexes that have been previously defined based on the period from January 2020 until September 2020. Based on this, the value for each index at the end of each month will be calculated and expressed as the index value for each month. Based on Figure 8, it illustrates how the analysis performed.

The analysis of the objective is about determining whether the Twitter platform can be used as a reliable and real-time source of information during the COVID-19 situation or not. The main reason for performing this analysis section relies on the components and functions of the HEALTH-EDRM (World Health, 2019), which states that the information should be the main focus, and this information should support the risk assessment function.

3.1.6 Data preprocessing

The Tweets related to the COVID-19 situation were retrieved. At the beginning of the data collecting process, tweets were collected by using the Twitter API by using the coding from Tweepy (Roesslein, 2015). These data were collected by the researcher from 21 January 2020 until 30 September 2020. The researcher set the criteria for collecting the related COVID-19 tweets by

1. The keywords for retrieving are “*Coronavirus*,” *nCoV*, “*COVID-19*,” where the term COVID-19 was added when WHO announced the new name for Coronavirus disease on 11 February 2020 (WHO 2019).
2. Each day the tweets were collected by the amount of 1,000 tweets.
3. The data were collected in the form of a text file.
4. The retweets data from Twitter were ignored due to its redundancy.

After finishing collecting the data, the next step is to perform the data preprocessing. This step for performing the data analysis is very crucial, and it is the fundamental step of the data analysis (S. Zhang, Zhang, & Yang, 2003). Based on the literature review part, this research includes mainly three steps for data preprocessing, which are

1. Tokenizing
2. Stop word removal
3. Stemming.

3.1.7 Content similarity analysis

After preparing for the process of data preparation, the next step of the research design based on the second objective of this research is to perform the analysis for content similarity. In this analysis, the tweets that have been previously performed the preprocessing function will be used to compare the similarity of the contents with the reliable data sources. The data from the WHO’s situation reports are the main source to be used to compare with the tweets. The WHO’s situation reports are the documents from the WHO that reports the daily COVID-19 situation. Based on this, the reports from 21 January 2020 were collected until 30 September 2020.

For the content similarity analysis, the content of tweets and the WHO’s situation reports will be compared based on different setting criteria according to Table 10. In Table 10, for the daily based, the one-day tweets will be compared with the data from the WHO’s situation report by using the same day for both of the data. The same setting will be applied to other criteria, but there is a difference in time setting. The content similarity analysis will be performed by using the Cosine similarity algorithm for the term-based similarity and. Therefore, the tweets and the WHO’s situation reports can be compared, and it can see whether tweets can be used as a reliable and real-time source of information or not. Based on this, for the initial phase of this objective, the content similarity analysis will be performed to prove whether the tweets data that have been collected have similar contents with the whole Twitter data for each day or not. In addition, the reliability of the Twitter account needs to be also checked to determine whether the tweets from that account are similar to the reliable data sources or not.

Table 10 Criteria setting for content similarity analysis

Criteria setting	Definition
Daily based comparison	The comparison between the tweets and WHO's situation report will be performed on day to day basis
Weekly based comparison	The comparison between the tweets and WHO's situation report will be performed on a week to week basis
Monthly based comparison	The comparison between the tweets and WHO's situation report will be performed on a month to month basis
Quarterly based comparison	The comparison between the tweets and WHO's situation report will be performed on a quarter to quarter basis

In conclusion, the research design based on the two objectives of this research can be concluded as Figure 9.

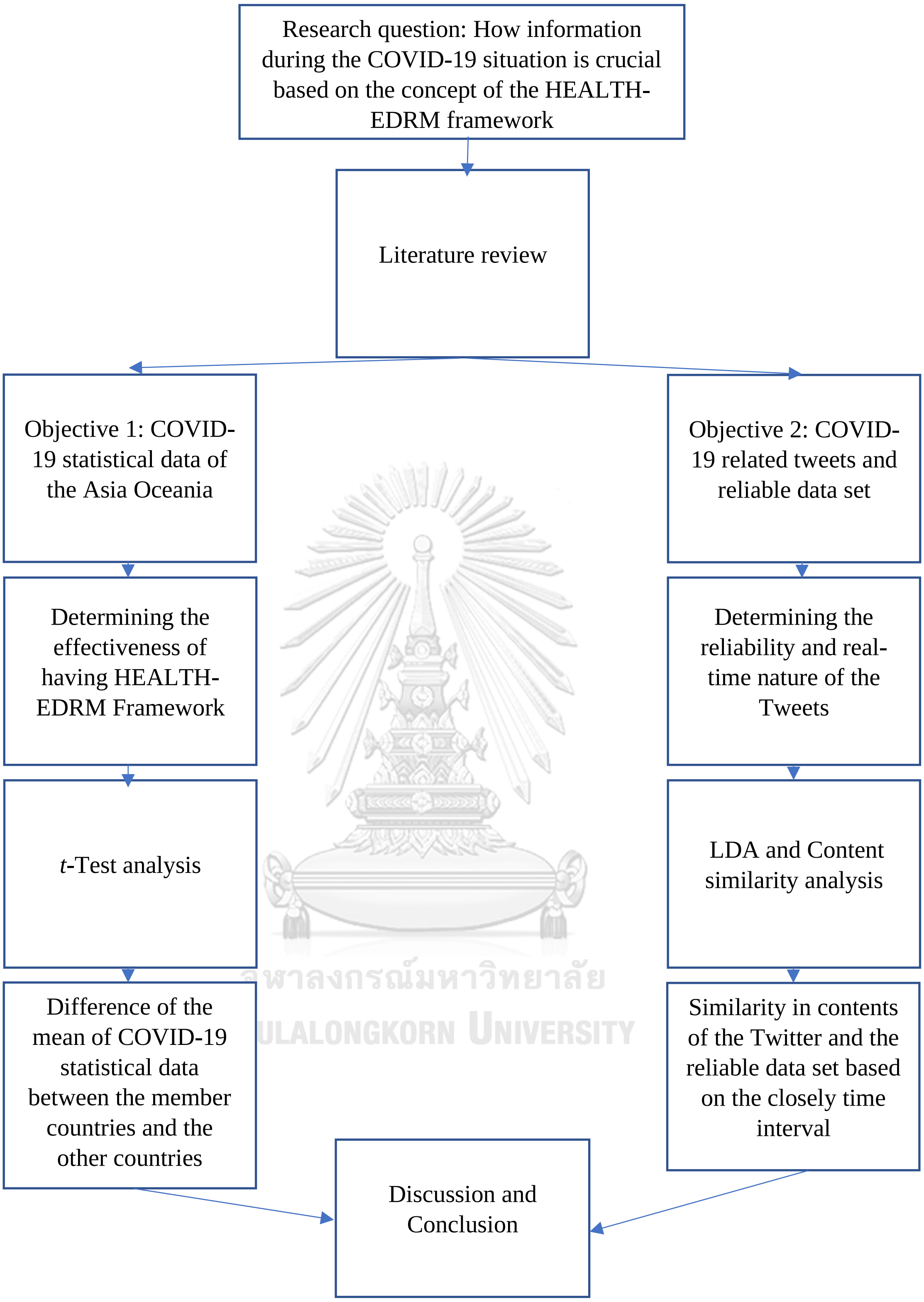

Research question: How information during the COVID-19 situation is crucial based on the concept of the HEALTH-EDRM framework
Literature review
Objective 1: COVID-19 statistical data of the Asia Oceania
Determining the effectiveness of having HEALTH-EDRM Framework
t-Test analysis
Difference of the mean of COVID-19 statistical data between the member countries and the other countries
Objective 2: COVID-19 related tweets and reliable data set
Determining the reliability and real-time nature of the Tweets
LDA and Content similarity analysis
Similarity in contents of the Twitter and the reliable data set based on the closely time interval
Discussion and Conclusion

Figure 9 Overall process of the research design

3.2 Research hypotheses

Based on the research design that has been previously explained, the research hypothesis of this research is separated into two parts based on the two objectives of this analysis.

3.2.1 Hypotheses for the first objective

The performance of the HEALTH-EDRM framework will be defined by the analysis of the impact of the COVID-19 situation for Asia Oceania countries and territories. The hypothesis of this experiment is separated into three main hypotheses.

First hypothesis: The APM for the HEALTH-EDRM members are lower than the HEALTH-EDRM non-members for either cross-section or continuous time series analysis.

Second hypothesis: The DPM for the member states of the HEALTH-EDRM are lower than the HEALTH-EDRM non-members for either cross-section or continuous time series analysis.

Third hypothesis: The RDA for the HEALTH-EDRM members are lower than the HEALTH-EDRM non-members for either cross-section or continuous time series analysis.

Based on this, the hypotheses of this research can be assured and fulfilled of the goal and vision of the HEALTH-EDRM framework, which is about building health resilience (Wright et al., 2020) and reduce the impact of multi kinds of disasters (World Health, 2019).

3.2.2 Hypotheses for the objective for the second objective

The determining of the nearly real-time and reliability of the Twitter microblogging platform will be determined by using content similarity analysis based on the Twitter text data. The hypothesis of this objective is divided into two hypotheses.

First hypothesis: The tweets can be used as a reliable source of data for providing the news and contents in the time of the COVID-19 situation.

Second hypothesis: The tweets can be used as a nearly real-time source of information during the time of the COVID-19 situation.

Based on these two hypotheses, if these two hypotheses are correct, it can be referred to Pourebrahim et al., which define the tweets can be used as real-time information based on the hurricane Sandy (Pourebrahim, Sultana, Edwards, Gochanour, & Mohanty, 2019). In addition, the hypothesis of this research also matches with Broersma and Graham, which define the content from Twitter as an effective source for searching for news and information.

3.3 Methodology

According to the literature review and research design part, two methodologies are presented in this section to fulfill the two objectives in this research.

3.3.1 COVID-19 statistical analysis

Based on the first research objective, the data from the two famous data sources are selected to be used in the analysis. These two data sources are ourworldindata.org (Roser et al., 2020) and worldmeter.info (Worldometers, 24 September, 2020). These two data sources provide a lot of useful data and information related to the COVID-19 situation. The examples of the data and information from these two data sources are accumulated of total affected cases for each country and territory, accumulated of death cases for each country, the accumulated number of test cases for each country, etc. Based on these sources of data, the analysis for checking the effectiveness of the HEALTH-EDRM framework will be performed by finding the difference of the accumulated number of affected and death cases for the country in the Asia Oceania region between the country that is the member states of this framework and the country that are not the member state of this framework.

t-Test analysis or Student's *t*-Test is selected to be the tool for this analysis. This analysis tool was initially developed by William Sealy Gosset in the year of 1908 (Haynes, 2013). The objective of the *t*-Test analysis is to check whether the two groups of the population have a difference in terms of population means or not (Haynes, 2013). The null hypothesis of the *t*-Test is stated as there is no difference between the two population means. The only two assumptions of the *t*-Test analysis are both of the two populations should have the same units of measurements. Another assumption for the *t*-Test analysis is to check whether the two population has the same variance or not. Based on this, for the case that both of the two population variances are assumed equal, equation (3.4) is the equation for finding the *t*-test result. Based on equation (3.4), μ_i is represented as the population average of population 1 and 2. n_i is represented as the number of data inside each population, and s_i is represented as the sample standard deviation of each population. In contrast, for the case that both variances of the two populations are assumed not equal, in this case, the Welch's *t*-Test will be used instead of the student's *t*-Test. The equation of the Welch *t*-test can be shown in equation (3.5). From the results of these two cases of the *t*-Test analysis, both of them need to compare the result with the critical value. The critical value of the student's *t*-Test can be found by the value of *t*-distribution with n_1-n_2-2 degree of freedom. On the other hand, the critical value of Welch's *t*-Test uses the value of t-distribution with the degree of freedom that calculates by equation (3.6) (Sakai ,2016). For the critical value of each case, there is another vital variable that involves finding the value in the *t*-Test distribution. This variable is called a significant level or α. This variable is the value of the probability that the analysis is correctly rejecting the null hypothesis. After obtaining the result from the *t*-Test analysis, both the critical value and test value from the calculation will be compared to each other. Based on this, if the value of the critical value is higher than the test value from the calculation, it will indicate that the null hypothesis shall not be ignored, or it can simply define that the means of both populations are the same. In contrast, if the critical value is less than the test value from the calculation, it will indicate that the null hypothesis should be ignored, or it can simply define that the means of both populations are not equal. Furthermore, there are also another application based on the *t*-test analysis which called paired *t*-test analysis. The main concept of the paired *t*-test is to compare the difference between the two dependent pairs of observations (Hsu & Lachenbruch, 2005). The difference between the dependent pair must be calculated and then the average of the difference between all pairs can be found. The test statistic can be found by the formula (3.7) (Hsu & Lachenbruch, 2005). In this case, the value of $\bar{d}$ is the value for average of the difference between all pair of the data. The value of the s_d is the sample standard deviation and the n is the value for the total amount of pairs. The assumption for the pair *t*-test is the same as the normal *t*-test. In this research, the impact data of the COVID-19 situation are collected at the end of each month. Therefore, in each month, the difference of the data can be calculated.

$$t = \frac{\mu_1-\mu_2}{\sqrt{{s_p}^2\frac{1}{n_1}+\frac{1}{n_2}}} \text{ where } s_p = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \qquad (3.4)$$

$$t = \frac{\mu_1-\mu_2}{\sqrt{\frac{s_1}{n_1}+\frac{s_2}{n_2}}} \qquad (3.5)$$

$$df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac{s_1}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2}{n_2}\right)^2}{n_2-1}\right)} \qquad (3.6)$$

$$t = \frac{\bar{d}}{{s_d}/{\sqrt{n}}} \qquad (3.7)$$

3.3.2 Latent Dirichlet Allocation (LDA)

From the second objective of this research, before determining the nearly real time and reliability of the Twitter data, the Latent Dirichlet Allocation or LDA have been selected to be the algorithm for generating the topic from both Twitter data and WHO's situation report. The main concept of the LDA algorithm is to generate the set of topics based on the related probability, which has been generated by the distribution of the words inside the corpus or documents (Blei, Ng, & Jordan, 2003). For the full explanations, they can be found in (Blei et al., 2003).

3.3.3 Contents similarity

For the second objective of this research, two groups of data were collected to perform the analysis. First, the tweets related to COVID-19 from the Twitter social media is selected to check whether it can be used in the nearly real-time or not. The tweets were collected by using the Twitter API by using the coding from Tweepy (Roesslein, 2015). These data were collected by the researcher from 21 January 2020 until 30 September 2020 for the amount of 1000 tweets per day. The query terms for collecting the data are “Coronavirus,” “nCov”, “COVID-19” where the term COVID-19 was added when WHO announced the new name for Coronavirus disease on 11 February 2020 (World Health, 2020). Second, the WHO’s situation reports were also collected from 21 January 2020, which is the day that the first situation report from WHO had been published (WHO 2019) until 30 September 2020. The content of the WHO’s situation reports were selected only the highlighted summary to capture all the events related to COVID-19 that has been happened all around the world, such as the world situation about the COVID-19, The country that has just announced the new cases for COVID-19, etc. The analysis is conducted by comparing the tweet's data together with the highlighted summary of the WHO’s situation.

In this analysis, the data from Twitter and WHO’s situation reports will be compared to check whether the data from Twitter can be used as reliable and real-time information during the COVID-19 situation or not. For the preprocessing stage, all of the data need to convert to the vector form based on the use of Term Frequency-Inverse Document Frequency (TF-IDF). The concept of TF-IDF is generated by using two terms, which are TF (Term Frequency) and IDF (Inverse Document Frequency). For the term TF, it is the relative frequency of the specific term (i) inside the specific corpus or document(d), which can be shown as equation (3.8). For the term IDF, it is the inverse of the relative frequency of the specific term (i) over the size of entire documents or corpora (Das, 1994), which can be shown by equation (3.9). After that, two terms are combined, the equation for TF-IDF can be shown as equation (3.10) (Ramos ,2003).

$$TF = f_{i,d} \tag{3.8}$$

$$IDF = log\frac{N}{f_{i,D}} \tag{3.9}$$

$$TF - IDF = f_{i,d} * log\frac{N}{f_{i,D}} \tag{3.10}$$

After converting all the data in text format into the vectors format, the process for finding the similarity between the two documents, the algorithm called Cosine Similarity, is selected for performing the similarity analysis. The concept of the Cosine Similarity is about converting each string to the high-dimensional vector space and compare the distance between the vector by calculating the dot product of the two high-dimensional vectors (Tata & Patel, 2007) (Liu et al. 2004). The formula of the cosine similarity can be expressed by equation (3.10) (Lahitani, Permanasari, and Setiawan ,2016), where A is the vector’s weight based on the text from documents A and B is the vector’s weight based on the text from document B. The tweets are represented as document A, and the data from the reliable source of information are represented as document B. The same study also indicated that the higher angle based on the two vectors indicates the high similarity index based on the two documents.

$$cos\ \alpha = \frac{A*B}{|A|*|B|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2 * \sum_{i=1}^{n}(B_i)^2}} \qquad (3.10)$$

Chapter 4 Results

In this chapter, the results of both objectives are shown. The results are shown in both qualitative and quantitative outcomes. In addition, this part also shows how the analysis that has been previously identified in chapter 3 can answer the purposes of this research.

4.1 Analysis for objective 1

4.1.1 Descriptive statistic for objective 1

In this objective, the study area is in the Asia Oceania region. In this region, the countries are separated into four main difference categories, which are Developing country (DC), Transition in the economic country (T), Developing countries (DPC), Least developing country (LDC). This type of country has been defined by United Nations Conference on Trade and Development (UNCTAD, 2020) and International Monetary Fund (IMF, 2000). Based on this, the country in Asia Oceania is also separated based on whether that country is a member of the HEALTH-EDRM (H) or Non-member of the HEALTH-EDRM (N). This information is summarized in Table 11 and Figures 10, and 11.

Table 11 Number of countries in Asia Oceania

	DC	T	DPC	LDC	Total (countries)
Non-Health EDRM Member	0	6	26	5	37
HEALTH-EDRM Member	3	0	11	3	17
Total (countries)	3	6	37	8	54

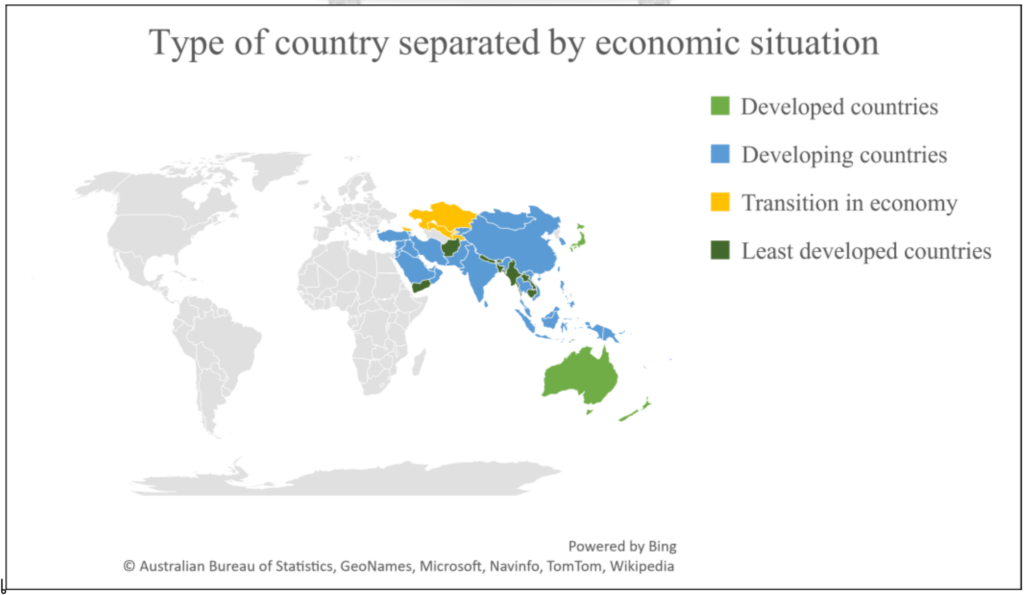

Type of country separated by economic situation
🟩 Developed countries
🟦 Developing countries
🟨 Transition in economy
🟩 Least developed countries
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Figure 10 The countries in Asia Oceania separated by economic situation

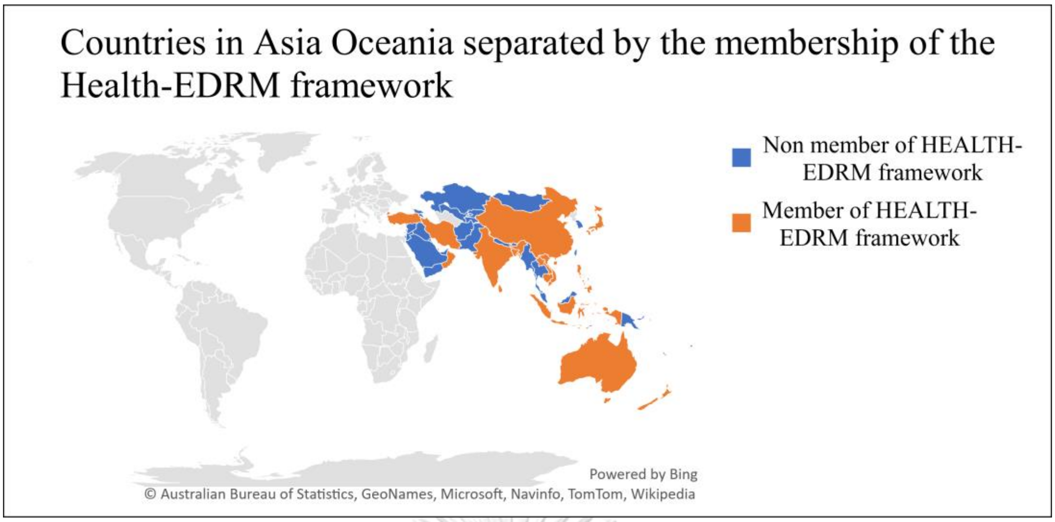

Countries in Asia Oceania separated by the membership of the Health-EDRM framework
🟦 Non member of HEALTH-EDRM framework
🟧 Member of HEALTH-EDRM framework
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Figure 11 The countries in Asia Oceania separated by the membership of the HEALTH-EDRM framework

The accumulated number of affected cases, death cases, and the ratio of death and affected cases are used for performing the analysis. The data that has been used for analysis in this research is collected from 31 December 2020 until 31 October 2021. Based on this, the total data related to the COVID-19 situation for ten months are used for performing the analysis. As of 31 October 2021, the descriptive statistic can be shown in Tables 12 and 13.

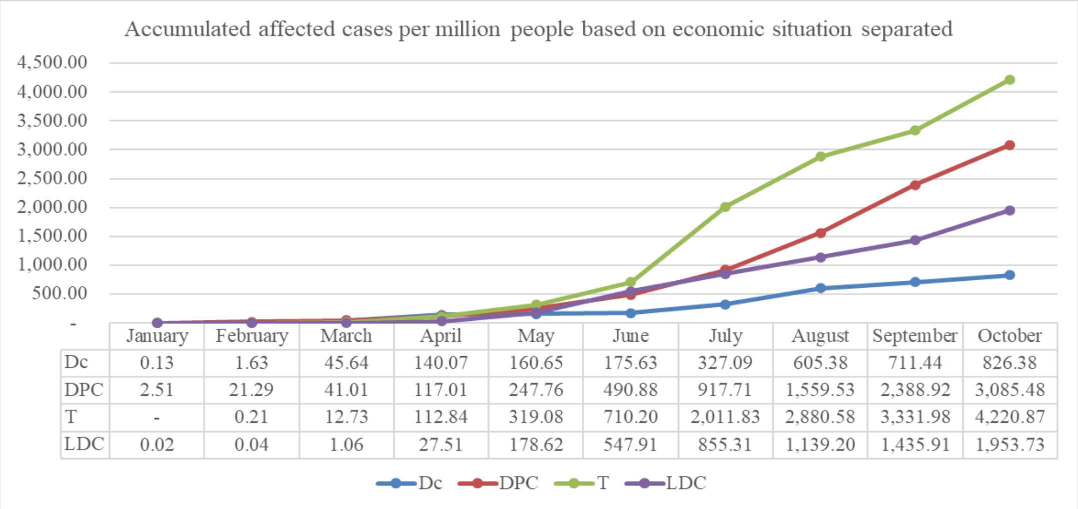
Table 12 Descriptive statistic for COVID-19 situation separated economic situation (As of 31 October 2020)

	Dc	T	DPC	LDC
Accumulated affected cases per million people (APM)	826.38	4,220.87	3,085.48	1,953.73
Accumulated death cases per million people (DPM)	17.14	51.32	55.57	29.67
Accumulated ratio of death and affected cases per million people (RDA)	0.02	0.01	0.02	0.02

Table 13 Descriptive statistic for COVID-19 situation separated member of the membership of HEALTH-EDRM framework (As of 31 October 2020)

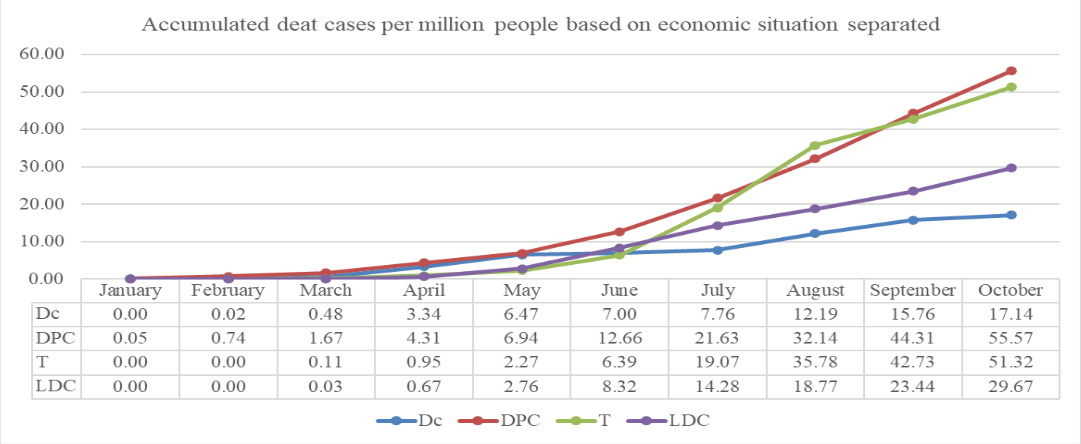
	Member of HEALTH-EDRM framework	Non-member of HEALTH-EDRM framework
Accumulated affected cases per million people (APM)	2,817.74	3,553.49
Accumulated death cases per million people (DPM)	52.51	51.25
Accumulated ratio of death and affected cases per million people (RDA)	0.01	0.01

Based on this, the data for the accumulated COVID-19 data can be treated as the time-series data by using the accumulated data on the last day of that month (e.g., the accumulated COVID-19 cases can on 31 January 2021 are represented as the time-series data for January). The time-series data can be shown in Figures 12, 13, 14, 15, 16, and 17.



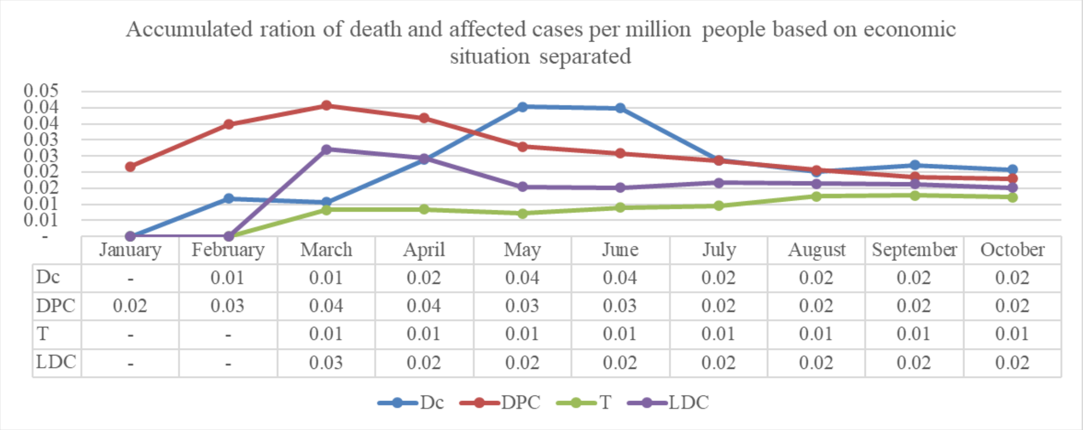
	January	February	March	April	May	June	July	August	September	October
Dc	0.13	1.63	45.64	140.07	160.65	175.63	327.09	605.38	711.44	826.38
DPC	2.51	21.29	41.01	117.01	247.76	490.88	917.71	1,559.53	2,388.92	3,085.48
T	-	0.21	12.73	112.84	319.08	710.20	2,011.83	2,880.58	3,331.98	4,220.87
LDC	0.02	0.04	1.06	27.51	178.62	547.91	855.31	1,139.20	1,435.91	1,953.73

Figure 12 The time-series data for COVID-19 affected cases separated by the economic situation



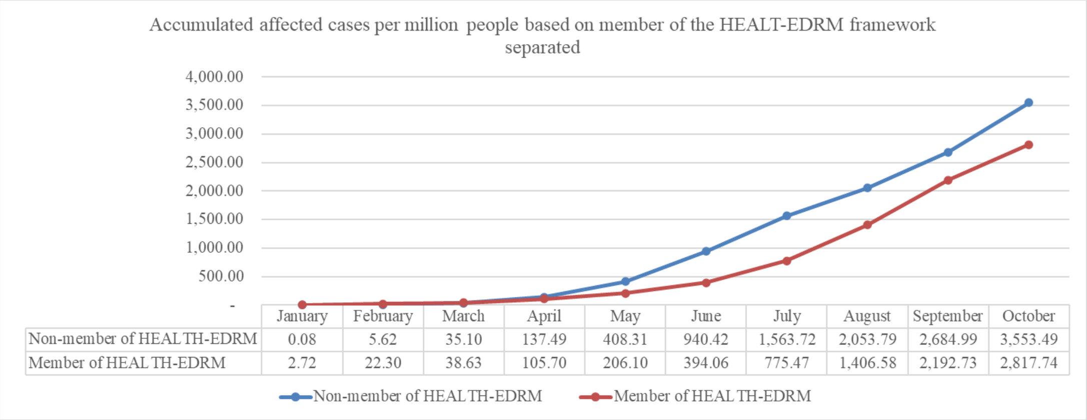
	January	February	March	April	May	June	July	August	September	October
Dc	0.00	0.02	0.48	3.34	6.47	7.00	7.76	12.19	15.76	17.14
DPC	0.05	0.74	1.67	4.31	6.94	12.66	21.63	32.14	44.31	55.57
T	0.00	0.00	0.11	0.95	2.27	6.39	19.07	35.78	42.73	51.32
LDC	0.00	0.00	0.03	0.67	2.76	8.32	14.28	18.77	23.44	29.67

Figure 13 The time-series data for COVID-19 death cases separated by the economic situation



	January	February	March	April	May	June	July	August	September	October
Dc	-	0.01	0.01	0.02	0.04	0.04	0.02	0.02	0.02	0.02
DPC	0.02	0.03	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.02
T	-	-	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LDC	-	-	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02

Figure 14 The time-series data for COVID-19 ratio of death and affected cases separated by the economic situation



	January	February	March	April	May	June	July	August	September	October
Non-member of HEALTH-EDRM	0.08	5.62	35.10	137.49	408.31	940.42	1,563.72	2,053.79	2,684.99	3,553.49
Member of HEALTH-EDRM	2.72	22.30	38.63	105.70	206.10	394.06	775.47	1,406.58	2,192.73	2,817.74

Figure 15 The time-series data for COVID-19 affected cases based on the member of the HEALTH-EDRM framework separated

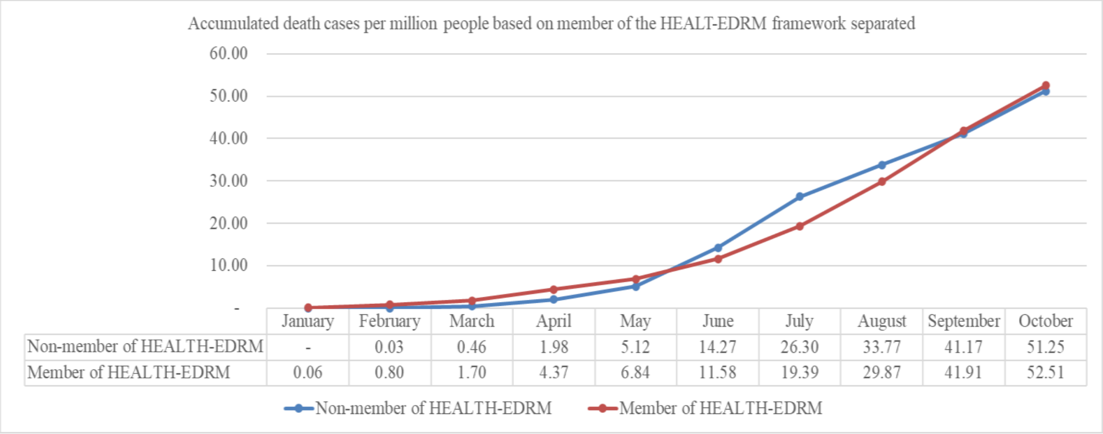


Figure 16 The time-series data for COVID-19 death cases based on the member of the HEALTH-EDRM framework separated

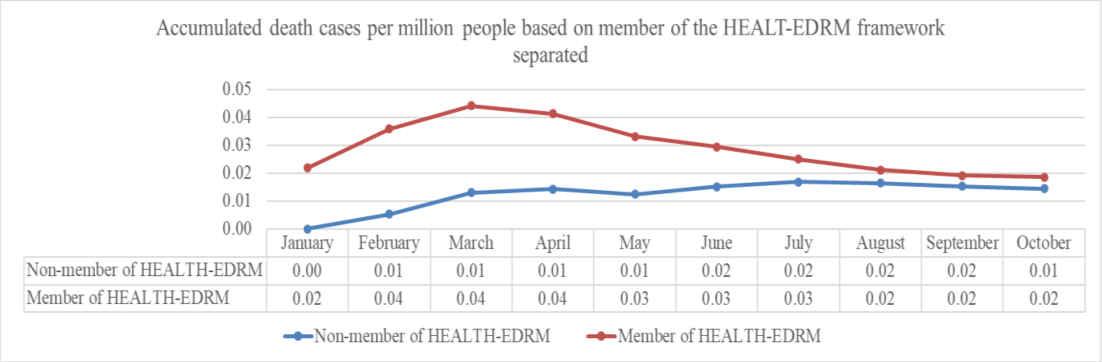


Figure 17 The time-series data for COVID-19 ratio of death and affected based on the member of the HEALTH-EDRM framework separated

4.1.2 Result of the analysis for objective 1

In this objective, the design of the experiment is set to be performed in 3 sets for four main aspects of hypotheses which can be defined as

Hypothesis 1: The effectiveness for dealing with COVID-19 of the **HEALTH-EDRM members** is better than the **HEALTH-EDRM non-member** countries (Overall)

Hypothesis 1.A: Affected cases per million (APM)
Hypothesis 1.D: Death cases per million (DPM)
Hypothesis 1.R: Ratio of death and affected cases (RDA)

Hypothesis 2: The effectiveness for dealing with COVID-19 of the **HEALTH-EDRM members** is better than the **HEALTH-EDRM non-member countries (Same economic situation)**

Hypothesis 2.A: Affected cases per million (APM)
Hypothesis 2.D: Death cases per million (DPM)
Hypothesis 2.R: Ratio of death and affected cases (RDA)

Hypothesis 3: The effectiveness for dealing with COVID-19 within **HEALTH-EDRM members for the better economic country** is better than **the worse economic country.**

Hypothesis 3.A: Affected cases per million (APM)

Hypothesis 3.D: Death cases per million (DPM)
Hypothesis 3.R: Ratio of death and affected cases (RDA)

Hypothesis 4: The effectiveness for dealing with COVID-19 within non-**HEALTH-EDRM members for the better economic country** is better than **the worse economic country**

Hypothesis 4.A: Affected cases per million (APM)
Hypothesis 4.D: Death cases per million (DPM)
Hypothesis 4.R: Ratio of death and affected cases (RDA).

In order to prove these hypotheses, the paired *t*-test analysis will be used as the tool for performing the analysis. The design of the analysis will be done by comparing the time-series data between the two countries. Therefore, the overall picture for each hypothesis can be summarized in Table 14.

Table 14 The analysis based on each hypothesis

Hypotheses	Comparison between
Hypothesis 1	H Vs. N
Hypothesis 2	H-Dpc Vs. N-Dpc
	H-Ldc Vs. N-Ldc
Hypothesis 3	H-Dc Vs. H-Dpc
	H-Dc Vs. H-Ldc
	H-Dpc Vs. H-Ldc
Hypothesis 4	N-Dpc Vs. N-T
	N-Dpc Vs. N-Ldc
	N-T Vs. N-Ldc

Based on this, the result of the paired one-tail *t*-test analysis can be shown in Tables 15, 16, 17.

Table 15 Result for paired t-test for COVID-19 affected cases

Hypotheses	Comparison between	Better performance	p-value
Hypothesis 1.A	H Vs. N	H	<0.05
Hypothesis 2.A	H-Dpc Vs. N-Dpc	H-Dpc	<0.05
	H-Ldc Vs. N-Ldc	N-Ldc	<0.05
Hypothesis 3.A	H-Dc Vs. H-Dpc	H-Dc	<0.05
	H-Dc Vs. H-Ldc	H-Dc	<0.05
	H-Dpc Vs. H-Ldc	H-Ldc	0.43
Hypothesis 4.A	N-T Vs. N-Dpc	N-Dpc	0.29
	N-Dpc Vs. N-Ldc	N-Ldc	<0.05
	N-T Vs. N-Ldc	N-Ldc	<0.05

Table 16 Result for paired t-test for COVID-19 death cases

Hypotheses	Comparison between	Better performance	p-value
Hypothesis 1.D	H Vs. N	H	0.29
Hypothesis 2.D	H-Dpc Vs. N-Dpc	H-Dpc	0.05
	H-Ldc Vs. N-Ldc	N-Ldc	<0.05
Hypothesis 3.D	H-Dc Vs. H-Dpc	H-Dc	<0.05
	H-Dc Vs. H-Ldc	H-Dc	<0.05
	H-Dpc Vs. H-Ldc	H-LDc	<0.05
Hypothesis 4.D	N-T Vs. N-Dpc	N-T	<0.05
	N-Dpc Vs. N-Ldc	N-Ldc	<0.05
	N-T Vs. N-Ldc	N-Ldc	<0.05

Table 17 Result for paired t-test for COVID-19 ratio of death and affected cases

Hypotheses	Comparison between	Better performance	p-value
Hypothesis 1.R	H Vs. N	N	<0.05
Hypothesis 2.R	H-Dpc Vs. N-Dpc	N-Dpc	<0.05
	H-Ldc Vs. N-Ldc	H-Ldc	<0.05
Hypothesis 3.R	H-Dc Vs. H-Dpc	H-Dc	0.05
	H-Dc Vs. H-Ldc	H-LDc	<0.05
	H-Dpc Vs. H-Ldc	H-Ldc	<0.05
Hypothesis 4.R	N-T Vs. N-Dpc	N-T	<0.05
	N-Dpc Vs. N-Ldc	N-Dpc	<0.05
	N-T Vs. N-Ldc	N-T	<0.05

From Tables 15, 16, and 17, these results show how the paired *t*-Test analysis has been performed based on the four hypotheses that have been defined. From this analysis, the group of countries with lower amounts of the APM, DPM, and RDA is shown together with the value of the *p*-value of the paired one-tailed *t*-test analysis.

Based on hypothesis 1, the result shows that the amounts of APM for the HEALTH-EDRM members are significantly lower than the HEALTH-EDRM non-members. However, the DPM for the HEALTH-EDRM members is not significantly lower than the HEALTH-EDRM non-members because the *p*-value is higher than 0.05. In contrast, the result for RDA shows that the amount of the RDA of the HEALTH-EDRM non-members is significantly lower than the HEALTH-EDRM members because the *p*-value is lower than 0.05.

From Hypothesis 2, based on the result of the *p*-value, the amount of the APM for the HEALTH-EDRM members, which are the developing country, is significantly lower than the HEALTH-EDRM non-members, which are also the developing country. In contrast, the amount of the APM for the HEALTH-EDRM members, which are the least developed country, is not significantly lower than the HEALTH-EDRM non-members, which are also the least developed country. The amount of DPM for both HEALTH-EDRM members, which are the developing country and least developed country, is not significantly lower than the HEALTH-EDRM members, which are also the developing country and least developed country. Lastly, only the amount of RDA for HEALTH-EDRM members, which are the least developed country, is significantly lower than the HEALTH-EDRM non-members, which are the least developed country.

From Hypothesis 3, based on the result of the *p*-value, the amounts of APM and DPM for the HEALTH-EDRM members, which are the developed countries, are significantly lower than both of the developing countries and least developed countries. However, the amounts of APM and DPM for the HEALTH-EDRM members, which are the developing country, are not significantly lower than the least developed countries. In addition, the amount of the RDA for the HEALTH-EDRM members, which have a better economic situation, is not significantly lower than the HEALTH-EDRM members, which have a worse economic situation.

From Hypothesis 4, based on the result of the *p*-value, the amount of the APM for the HEALTH-EDRM non-members, which have a better economic situation, is not significantly lower than the HEALTH-EDRM non-members, which have a worse economic situation. Similarly, only the amount of DPM for the HEALTH-EDRM non-members, which are the transition in economic, is significantly lower than the HEALTH-EDRM non-members, which are the developed country. In contrast, based on the value of the RDA, the results show that the better economic situation country, which is the HEALTH-EDRM non-members, have significantly lowered the amount of RDA than the worse economic situation country which are the HEALTH-EDRM non-members.

Next, the summarized result of the paired *t*-test analysis of this result can be illustrated by Tables 18, 19, 20.

Table 18 Summarized result for paired t-test for COVID-19 affected cases

Hypotheses	Comparison between	Supported/Not supported
Hypothesis 1.A	H Vs. N	Supported
Hypothesis 2.A	H-Dpc Vs. N-Dpc	Supported
	H-Ldc Vs. N-Ldc	Not supported
Hypothesis 3.A	H-Dc Vs. H-Dpc	Supported
	H-Dc Vs. H-Ldc	Supported
	H-Dpc Vs. H-Ldc	Not supported
Hypothesis 4.A	N-Dpc Vs. N-T	Not supported
	N-Dpc Vs. N-Ldc	Not supported
	N-T Vs. N-Ldc	Not supported

Table 19 Summarized result for paired t-test for COVID-19 death cases

Hypotheses	Comparison between	Supported/Not supported
Hypothesis 1.D	H Vs. N	Not supported
Hypothesis 2.D	H-Dpc Vs. N-Dpc	Not supported
	H-Ldc Vs. N-Ldc	Not supported
Hypothesis 3.D	H-Dc Vs. H-Dpc	Supported
	H-Dc Vs. H-Ldc	Supported
	H-Dpc Vs. H-Ldc	Not supported
Hypothesis 4.D	N-Dpc Vs. N-T	Supported
	N-Dpc Vs. N-Ldc	Not supported
	N-T Vs. N-Ldc	Not supported

Table 20 Summarized result for paired t-test for COVID-19 ratio of death and affected cases

Hypotheses	Comparison between	Supported/Not supported
Hypothesis 1.R	H Vs. N	Not supported
Hypothesis 2.R	H-Dpc Vs. N-Dpc	Not supported
	H-Ldc Vs. N-Ldc	Supported
Hypothesis 3.R	H-Dc Vs. H-Dpc	Not supported
	H-Dc Vs. H-Ldc	Not supported
	H-Dpc Vs. H-Ldc	Not supported
Hypothesis 4.R	N-Dpc Vs. N-T	Supported
	N-Dpc Vs. N-Ldc	Supported
	N-T Vs. N-Ldc	Supported

4.2 Analysis for objective 2

4.2.1 Descriptive statistic for objective 2

In this objective, the comparison between the content related to the COVID-19 situation of the WHO's situation report and the Twitter data will be performed. The data of both WHO's situation report and the Twitter data had been collected from 21 January 2021 until 16 August 2021. The reason for selecting this time frame is that it is the first day that WHO announces the first situation report and the last day that WHO announces the daily situation report and changed it to be weekly situation report. Based on this, the criteria for collecting have been shown in Table 21.

Table 21 Criteria for collecting the data

	Twitter data	WHO's situation report
Number of daily contents	First 1,000 tweets in that specific day	Only the highlight content in that specific day
Keywords	Coronavirus, COVID, nCoV	All content in the highlighted content
Time difference of the data	GMT+0 (UTC)	GMT-1
Number of days for collecting the data	209 days (from 21 January 2021)	209 days (from 21 January 2021)

In this case, the researcher conducts some preliminary experiments to illustrate the character of the data set for both WHO and Twitter data by using the LDA algorithm. The parameters setting for the LDA algorithm are the number of topics generated and the number of key terms generated in one topic, learning offset, the number of iterations, which can be illustrated by Table 22. Based on this, the parameter setting for Table 22 is found based on trial and error and the reasonable results that come out from the LDA algorithm.

Table 22 The parameter setting for LDA algorithm

Parameter	Value
Learning offset	1
Number of iterations	1000
Number of topics	50
Number of key terms generated in one topic	100

4.2.2 Result of the analysis for objective 2

Next, the content similarity analysis will be performed to fulfill the objective of this research about identifying whether the Twitter data can be used as a reliable and nearly-real time source of information or not. The content similarity analysis will be performed by comparing the top topic, which generated the LDA algorithm, from the side of WHO's situation report and the Twitter data. In this case, the analysis is divided into four main time frames, which are daily based, weekly based, monthly based, and quarterly based.

For the daily based analysis, the experiment has been separated into three main categories. The first category is the comparison between the WHO data and Twitter data within the same day (T[1]&T). The second category is the comparison between the data of WHO one day before the Twitter data and the Twitter data in that day (T&T+1). For example, if the WHO data is the data on 23 January 2020, the Twitter data will be the data on 24 January 2020. The third category is the comparison between the data of WHO one day after the Twitter data and the Twitter data on that day (T+1&T). For example, if the WHO data is the data on 23 January 2020, the Twitter data will be the data on 22 January 2020. The results of the monthly average value of cosine similarity result of these daily based experiment are shown in Table 23.

Table 23 Result of the daily based experiment

	T&T		T&T+1		T+1&T	
	Average	SD	Average	SD	Average	SD
January	0.167	0.001	0.169	0.000	0.172	0.001
February	0.144	0.002	0.143	0.001	0.147	0.002
March	0.107	0.001	0.105	0.001	0.106	0.001
April	0.109	0.001	0.111	0.001	0.111	0.001
May	0.097	0.001	0.097	0.001	0.096	0.001
June	0.094	0.001	0.096	0.001	0.098	0.001
July	0.089	0.001	0.086	0.001	0.086	0.001
August	0.086	0.001	0.089	0.001	0.088	0.001

In order to clarify the daily based experiment, the maximum value of the daily based result for each month has been selected to be the representative value for each month, and it can be shown in Table 24.

Table 24 Monthly maximum value for the cosine similarity based on the daily based result

	T&T	T&T+1	T+1&T
January	0.210	**0.236**	0.205
February	0.225	**0.226**	0.212
March	0.151	0.164	**0.166**
April	0.146	0.168	**0.177**
May	**0.158**	0.154	0.149
June	0.167	0.170	**0.181**
July	0.139	0.138	**0.142**
August	0.139	0.163	**0.205**

[1] The first T is the data from WHO and the second T is the data from Twitter.

Based on this, in order to prove that the content from the Twitter data can be used for the reliable data source, the real data from the maximum value of Table 25 can be used to compare with the data from the reliable source of information. The news from the NBC news has been used to compare. The data from NBC news have been used in many studies, such as the study of Chen et al. (Chen, Lerman, & Ferrara, 2020b) used NBC news as the source of information to show the timeline of the COVID-19 situation, etc.

Table 25 Content from NBC based on the maximum value of the cosine similarity from daily based

Month	Date	Content
January	30 January 2020	WHO declared the outbreak a global public health emergency as more than 9,000 cases were reported worldwide, including in 18 countries beyond China.
February	13 February 2020	The first coronavirus death was recorded outside Asia. The patient was an 80-year-old Chinese tourist who died in France.
March	18 March 2020	The WHO announced an international trial to gather data about, which treatments are most effective for the coronavirus. Participants in the so-called solidarity trial include Argentina, Canada, France, Norway, South Africa, Spain, Switzerland and Thailand.
April	10 April 2020	The global coronavirus death toll crossed 100,000, according to a tally compiled by Johns Hopkins University.
May	19 May 2020	Italy reported fewer than 100 coronavirus deaths in a 24 hour period for the first time in nearly 10 weeks.
June	28 June 2020	Global death toll from COVID-19 surpassed 500,000 and the number of confirmed cases worldwide topped 10 million.
July	13 July 2020	-
August	6 August 2020	-

Based on the result of Table 25, it can be shown that the content in the day that has a high value of cosine similarity have the content that has a lot of impacts related to the COVID-19 situation, such as the number of world COVID-19 affected cases, the worldwide situation of the COVID-19 that have been shifted from Asia to other continents, and etc. In addition, the experiment for comparing the cosine similarity between the WHO data and the content from the news from the NBC news is also performed in order to find the expected value of the result of the cosine similarity when comparing the contents of the Twitter and WHO's situation report. The result of the average monthly cosine similarity for the period of January until March[2] is around 0.145.

Next, the analysis for the weekly, monthly, and quarterly have been performed by using the same methodology as the daily based analysis, and it can be shown in Tables 26, 27, 28.

[2] The cosine similarity experiment for finding the content similarity have to perform only the period of January until March due to the limitation amount of news in the later month.

Table 26 Weekly based cosine similarity result

Week [3]	Cosine similarity
Week 1	0.216
Week 2	0.214
Week 3	0.221
Week 4	0.225
Week 5	0.204
Week 6	0.164
Week 7	0.111
Week 8	0.126
Week 9	0.162
Week 10	0.144
Week 11	0.162
Week 12	0.206
Week 13	0.133
Week 14	0.142
Week 15	0.139
Week 16	0.127
Week 17	0.113
Week 18	0.146
Week 19	0.171
Week 20	0.135
Week 21	0.165
Week 22	0.155
Week 23	0.121
Week 24	0.129
Week 25	0.126
Week 26	0.114
Week 27	0.123
Week 28	0.105
Week 29	0.134
Week 30	0.124

[3] The 1st week started from 21 January 2020 to 27 January 2020 and continue based on the same pattern until week 25th which started from

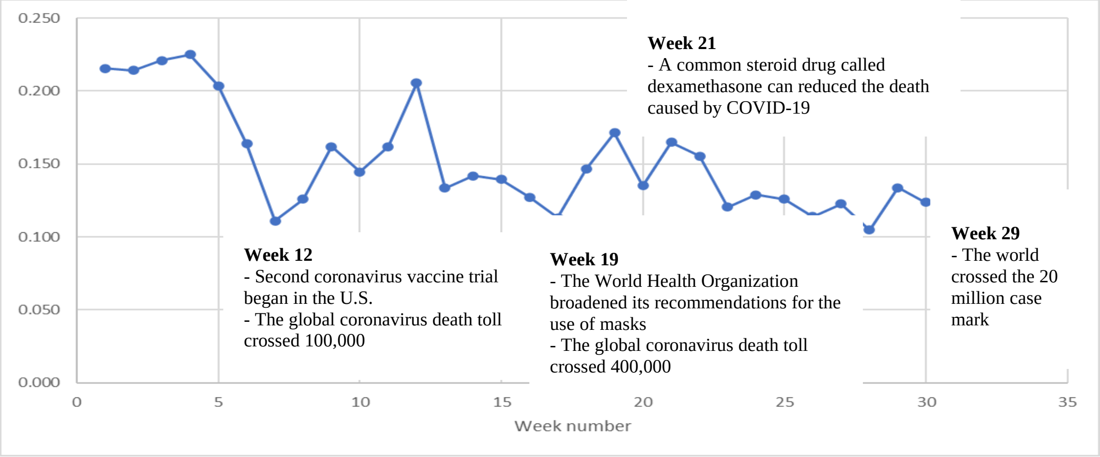
Table 27 Monthly based cosine similarity result

Month [4]	Cosine similarity
January	0.208
February	0.221
March	0.178
April	0.182
May	0.157
June	0.173
July	0.170
August	0.130

Table 28 Quarterly based cosine similarity result

Quarter	Cosine similarity
Quarter 1 (January-March 2020)	0.214
Quarter 2 (April-June 2020)	0.181
Quarter 3 (July-16 August 2020)	0.141

From the result of Tables 26, 27, and 28, it can be shown that the similarity of the content between the WHO and Twitter data are mostly similar at the beginning period of the COVID-19 situation, and the values tend to drop and then move up a bit if there are some important events happened in that period of time. Based on this, the result can be shown in Figures 18 and 19



Week 21
- A common steroid drug called dexamethasone can reduced the death caused by COVID-19

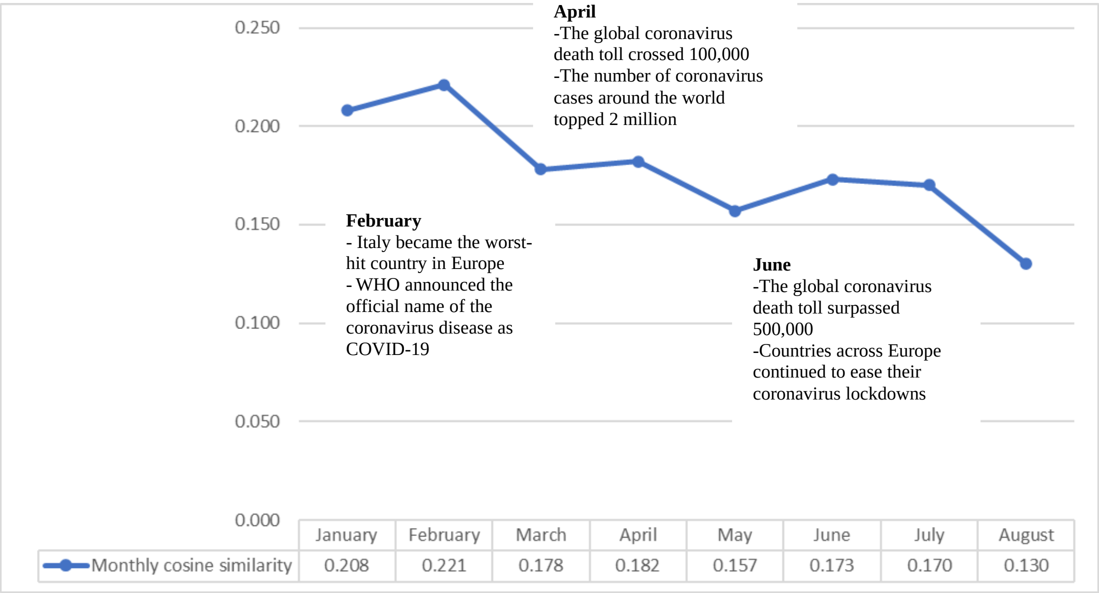
Week 12
- Second coronavirus vaccine trial began in the U.S.
- The global coronavirus death toll crossed 100,000

Week 19
- The World Health Organization broadened its recommendations for the use of masks
- The global coronavirus death toll crossed 400,000

Week 29
- The world crossed the 20 million case mark

Figure 18 Cosine similarity result (Weekly based)

[4] The cosine similarity for January is calculated based on the data from both WHO and Twitter from 21 January 2020 until 31 January 2020. In addition, the cosine similarity for August is calculated based on the data from both WHO and Twitter from 1 August 2020 until 16 August2020.



April
-The global coronavirus death toll crossed 100,000
-The number of coronavirus cases around the world topped 2 million

February
- Italy became the worst-hit country in Europe
- WHO announced the official name of the coronavirus disease as COVID-19

June
-The global coronavirus death toll surpassed 500,000
-Countries across Europe continued to ease their coronavirus lockdowns

	January	February	March	April	May	June	July	August
Monthly cosine similarity	0.208	0.221	0.178	0.182	0.157	0.173	0.170	0.130

Figure 19 Cosine similarity result (Monthly based)

From Figures 18 and 19, the trend of the cosine similarity based on monthly and weekly based have been shown. These two figures illustrate how the trend of the content similarity between the WHO data and Twitter drop from the beginning until the end of the study time frame. However, some of the big events or situations trigger the value of the cosine similarity during the study period.

Chapter 5 Discussion

In this chapter, the discussion for this research based on the two objectives is explained and clarified based on the results that have been shown in chapter 4.

5.1 Discussion for objective 1

For the first objective, based on the result that has been shown in Tables 15, 16, and 17, only the result based on the amount of the APM the member country shows the better performance to cope with the COVID-19 situation than the non-member country. The results from the first hypothesis can be acceptable because the HEALTH-EDRM framework is the framework that has been established since the end of 2019. Therefore, the member countries need to spend more time to fully apply the HEALTH-EDRM framework into their own disaster management system because the main goal of this framework is to make good cooperation between the disaster-related stakeholders such as the healthcare system disaster monitoring organization, etc. From the result of Hypothesis 2, the result of the APM and DPM shows that the HEALTH-EDRM framework can effectively help the developing country to deal with the COVID-19 situation. In addition, the result shows that the RDA for the least developed countries, which are the HEALTH-EDRM member, are better than the HEALTH-EDRM non-members. Based on this, the result can prove that the HEALTH-EDRM can help its members to manage the COVID-19 situation. In addition, from Hypothesis 3, for the value of APM and DPM, the developed countries show significantly better performance for dealing with the impact of the COVID-19 situation, which can be proved by the better economic and healthcare system. However, the result for APM, DPM, and RDA shows that the HEALTH-EDRM members, which are the least developed countries, have better performance for dealing with the COVID-19 situation than the HEALTH-EDRM members, which are the developing countries. This result is contradicted with the result of the Dünser et al. (Dünser, Baelani, & Ganbold, 2006) , which indicates that the low-income countries or the least developed countries still need more improvement for the healthcare system when comparing to the high-income country such as developed countries. In contrast, the result of this research can be used to reassure the result of Chan et al. (Chan & Shaw, 2020), which states that the objective and the usefulness of the HEALTH-EDRM framework is about cost-effectiveness. Moreover, the HEALTH-EDRM framework has a lot of effectiveness to deal with specific kinds of disasters, such as epidemiological, when applying to the low-income country or the least developed countries. Furthermore, based on the results of hypothesis 4, the amount of RDA of the non-member countries, which have a better economic situation, will better deal with the COVID-19 than the countries with a worse economic situation. Therefore, it indicates that the effectiveness for managing with the COVID-19 of the HEALTH-EDRM non-members depends purely on the economic situation. However, the results of the APM and DPM shows the different result from the amount of the RDA. They define that some of the HEALTH-EDRM non-members, which have the worse economic situation, have better effectiveness for dealing with the COVID-19 situation. Based on this, the further analysis can be performed for by analyzing others factors that might affect the impact of the COVID-19 such as the number of the tourists and the number of the foreigner.

5.2 Discussion for objective 2

For the second objective of this research, the result of the content similarity analysis shows that the content similarity between the highlight of the WHO's situation reports, and the Twitter data related to the COVID-19 situation tend to have more similarity in the beginning phase of the COVID-19 situation, which can be illustrated by the daily, weekly, monthly and quarterly based analysis. The reasons behind this result are that at the beginning phase of the COVID-19 situation, both of the data have the content related to the effects of the COVID-19, and the situation of the COVID-19, which can be shown by the terms affected cases, death, etc. In addition, the current COVID-19 situation in some specific countries or cities has been mostly mentioned based on the LDA results. In contrast, at the later phase of the COVID-19 situation (in quarter 3 of 2020), the result from the LDA show that the contents of the WHO's situation report are related to the operation from the WHO's officer together with some guidelines and policies for dealing with the COVID-19. However, the content based on the Twitter data tends to be more about the effects of the COVID-19 situation, such as the shutdown of some businesses, the announcement from the governments, and etc.

Based on the purpose of this objective about the nearly-real time manner of the Twitter data, from the result of the maximum daily cosine similarity for each month, it shows that the similarity between the WHO's situation report (T+1) and the Twitter data (T) tend to have more similarity at the beginning phase of the COVID-19 situation (January and February 2020), based on this result it can clarify that the information from the Twitter is faster than the information at the beginning phase of the COVID-19 situation based on this result can be matched with the study of the Kireyev et al. (Kireyev, Palen, and Anderson ., 2009), which state that the Twitter is the real-time and update source of information based on the disaster situation. In contrast, the content of the WHO's situation reports need to be summarized based on the news and information from various sources of information such as the CDC, the information from the health ministry from each country(WHO, 2020) before creating the daily situation report for that specific day.

For the purpose of this objective related to the reliability of the Twitter data, the result can be interpreted by the value of the cosine similarity. The cosine similarity result is based on both weekly and monthly based. It shows the downtrend of the value of the cosine similarity from the beginning until the end of the study period. The trends of the content similarity for all time frames are in the decline stage due to the significant decrease of the value of the cosine similarity. Based on this, the value of cosine similarity from the first week is decreased by 42.59% when compared with the last week of the analysis process. Similar to the result of the weekly based analysis, the value of the cosine similarity from the first period based on both monthly and quarterly based is decreased by 37.5% and 34.12% respectively when compared to the last period based on both monthly and quarterly based. However, from the Figures 18 and 19, there is still some period that the value of the cosine similarity increases and then drops. Based on this, if there are some important as stated in these Figures, the value of the cosine similarity will be increased. Therefore, the content based on the Twitter data is reliable based on the fact that the value of the cosine similarity between the WHO's situation reports, and the Twitter data increases when there are some important events that occur in that specific period of time. The results of the second hypothesis also match with the result of Rufai et al. (Rufai & Bunce, 2020), which state that at the beginning phase of the COVID-19, around 80 % of the content of Twitter were classified as informative information about the COVID-19 situation, so the value of the cosine similarity is high during the beginning phase of the COVID-19 due to the similarity of the informative information between both Twitter and WHO. The problem about the misinformation is also other factors that impact the value of the cosine. The study of Shahi et al. (Shahi, Dirkson, & Majchrzak, 2021) states that the amount of misinformation starts to increase since April 2020, which is almost the same period that the monthly cosine similarity results start to drop. Therefore, the information from Twitter can be more reliable during the beginning phase of the COVID-19. Furthermore, according to the result of the LDA, some of the content of Twitter in the later phase of the COVID-19 situation are mostly about the politic situation and also how well the government dealing with the COVID-19 situation.

Chapter 6 Conclusion

In this chapter, the conclusion of the research is explained in this chapter. In addition, the contribution, limitation, and possible future works are also explained and discussed in this chapter too.

6.1 Conclusion

The two objectives based on the COVID-19 situation have been concluded. For the first objective, the purpose of this objective is to find whether the HEALTH-EDRM framework can create effectiveness in order to deal with the COVID-19 situation or not. The paired one-tailed *t*-Test analysis is selected to be applied to perform the analysis. The data that have been selected are the amount of APM, DPM, and RDA from 1 January 2020 until 31 October 2020 for the countries in the Asia and Oceania continents. Based on this, the studied countries are divided into two main categories, which are the HEALTH-EDRM framework separated countries and economic separated countries. The analysis has been separated into four main hypotheses. The conclusion of these four hypotheses can be summarized into four main aspects. First, the HEALTH-EDRM members have more effective for dealing with the COVID-19 situation than the HEALTH-EDRM non-members. Second, the HEALTH-EDRM members have more effective for dealing with the COVID-19 situation than the HEALTH-EDRM non-members with the same economic separated situation. Third, the better economic separated countries, which are the HEALTH-EDRM members, have more effective for dealing with the COVID-19 situation than the worse economic separated countries, which are also the HEALTH-EDRM members. Lastly, the better economic separated countries, which are the HEALTH-EDRM non-members, have more effective for dealing with the COVID-19 situation than the worse economic separated countries, which are also the HEALTH-EDRM non-members. Based on this, the result shows that the HEALTH-EDRM members can only perform better in dealing with the COVID-19 situation than the non-member countries based on the amount of APM. However, the interesting result is that the result for Hypothesis 3 shows that the least developed countries, which are the HEALTH-EDRM members, can perform better in terms of dealing with the COVID-19 situation than the developing countries, which are considered as the better economic countries. In addition, the results of the last hypothesis regarding the value of the RDA show that the effectiveness of dealing with the COVID-19 is purely based on the economic situation for the non-member of the HEALTH-EDRM framework. However, further analysis needs to be performed.

For the second objective, the objective of the second objective is to check whether Twitter has nearly real-time properties and can be used as a reliable source of information or not. The LDA algorithm, together with the Cosine similarity index, has been selected to perform the analysis. The data that have been selected to be used to perform this analysis are the highlight of the WHO’s situation report and the Twitter data from the first day that the WHO announces their first daily situation report until the last day that WHO announces their last daily situation report. The methodology of this objective is to find the content similarity by using the cosine similarity index based on the topic that has been generated by the LDA algorithm. The analysis is separated into four different time frames, which are daily, weekly, monthly, and quarterly based. The result of the analysis based on the daily based result shown that in the first two months of the COVID-19 situation (January and February 2020), the content similarity between the WHO's situation reports on that specific day and the Twitter data in the day before the announcement of the WHO's situation report(T+1&T) tends to have more value of the cosine similarity than the content similarity between the same day of the WHO's situation report and the Twitter data (T&T), and also the content similarity between the WHO's situation report one day before the Twitter data and the content from the Twitter data in that specific day (T&T+1). Therefore, it can be proved that the content of the Twitter data can be used as a nearly real-time source of information. In addition, for the reliability property of the Twitter data, based on the result of the weekly and monthly based analysis, the trend of the cosine similarity of the comparison between the WHO's situation reports, and Twitter data has the downtrend. However, some week or some month of the trend of the cosine similarity will go up when there is some major or big event that happened all around the world. Based on this, it can be proved by using the NBC news, which is the other reliable information for the COVID-19 related news, to state the news that happened in that specific week or month. In addition, Twitter is considered a reliable source of information in the beginning phase of the COVID-19 situation due to the high value of the cosine similarity in the beginning phase of the COVID-19 situation.

In conclusion, the results from objectives 1 and 2 can be used to confirm the good disaster management can reduce the impacts from the disaster, such as COVID-19, which is one of the most severe disasters. Based on this, good disaster management requires some disaster management framework, such as the HEALTH-EDRM framework. Although the HEALTH-EDRM framework did not perform high performance dealing with the COVID-19 situation, the HEALTH-EDRM framework can be benefited to the country, which has low-income (Least developed countries). In addition, based on the objective of the HEALTH-EDRM framework, which requires the collaboration of the disaster-related organization, real-time information is very crucial for understanding the current situation of the disasters. Therefore, Twitter's social media platform is proved that it can be used as a nearly real-time and reliable source of information during the COVID-19 situation, especially in the beginning phase of the COVID-19 situation.

6.2 Contribution

The result from both objectives can be proposed and applied in many related fields. For the contribution of the first objective, the discussion from the first objective stated that the HEALTH-EDRM framework could be proposed to the low-income country in order to apply based on their cost-effectiveness characteristic of this framework, which can be helped to reduce the amount of affected and death people.

For the second objective, the discussion of the second objective stated that Twitter could also be used as the alternative source of information for the time of the disaster due to their nearly real-time manner of Twitter. In addition, Twitter can be suggested as the other source of news and information during the time of the disaster for both affected people and the organization who have been in charge of dealing with the disaster.

6.3 Limitation

Both objectives still have some limitations for the scopes and the analysis methods. For the first objective, there are four main limitations. First, the dimension of the data for coping with the COVID-19 situation is focusing only on the affected cases and death cases due to the limitation and incompleteness of the data. Second, the study period included in this research is not enough when comparing to the present situation because currently, some countries still have the COVID-19 situation for their second and third wave already. Third, the scope area of this study might not fulfill the affected area all around the world due to the spreading of the COVID-19 that has been affected the world since the beginning of 2020. Lastly, the other factors that might be affected the COVID-19 situation, such as the number of tourists, the degrees of the policy that the country used, are not included in this research.

For the limitation of the second objective, there are three main limitations that need to be concerned. First, the amount of the data for both Twitter and reliable data source are still less to perform the bigger scope of the research, so the amount of the data from both Twitter and the reliable data source should be increased especially for the amount of the twitter data which is only 1,000 tweets per day. However, the problem about the fewer amount of data is preliminary solved by excluded the retweets, which are considered as redundant information. Second, the irrelevant terms from both WHO data and the Twitter data still do not exclude before performing the analysis. Therefore, there might have some effect when calculating the cosine similarity index. Lastly, the content about the COVID-19 vaccine is not included in this research due to the short study period. Based on this, the result of the cosine similarity might be increased to be the same level as the initial phase of the COVID-19 when people are mostly focused on the impact of the COVID-19.

6.4 Future work

For the future work of this research, there are still a lot of rooms to be improved based on the two main objectives of this research. For the first objective, the number of factors related to the COVID-19 situation, such as the intensity of the rule and regulation, the amount of COVID-19 test cases, can be used for extending the scope of the analysis. In addition, the extension of the study area and period can be done for observing the change of the trend based on the COVID-19 impacts because there is no clear evidence states that when the COVID-19 situation will not impact our world. Furthermore, if there is evidence stated that the HEALTH-EDRM members are fully applied this framework in their country, the effectiveness of this HEALTH-EDRM framework can be the motivation for the other members, and it can also be proposed to the non-member countries to apply the HEALTH-EDRM framework for their country too.

For the second objective, there are also rooms for improvement by using the classification algorithm such as the concept of Support vector machine (SVM), which is one of the most robust machines learning techniques (Subasi, Kevric, & Canbaz, 2019) in order to group the content from both of the WHO’s situation report and the Twitter data into the specific topic based on this the accuracy for measuring the similarity between the content will be increased. In addition, the study of the Monti et al. (Monti, Frasca, Eynard, Mannion, & Bronstein, 2019) shows that the Convolutional Neural Network or CNN can be applied for detecting the contents that considered as fake news. Therefore, the result of the cosine similarity can be improved based on the reduction of the irrelevant data.

REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). An introduction to text mining. In *Mining text data* (pp. 1-10): Springer.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*(1), 45-65.

Al-Dahash, H., Thayaparan, M., & Kulatunga, U. (2016). *Understanding the terminologies: Disaster, crisis and emergency*.

Baldwin, R., & Weder di Mauro, B. (2020). Economics in the Time of COVID-19. In: CEPR Press.

Bär, D., Biemann, C., Gurevych, I., & Zesch, T. (2012). *Ukp: Computing semantic textual similarity by combining multiple content similarity measures*.

Bird, S., Klein, E., & Loper, E. (2008). NLTK Documentation. *Online: accessed April*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Boyd, D., Golder, S., & Lotan, G. (2010). *Tweet, tweet, retweet: Conversational aspects of retweeting on twitter*.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., . . . Grobler, J. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Buneman, P., Davidson, S., & Suciu, D. (1995). *Programming constructs for unstructured data*.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine, 9*(2), 48-57.

Chae, J., Thom, D., Jang, Y., Kim, S., Ertl, T., & Ebert, D. S. (2014). Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics, 38*, 51-60.

Chan, E. Y. Y., & Shaw, R. (2020). *Public Health and Disasters*: Springer.

Chatfield, A. T., & Brajawidagda, U. (2013). *Twitter early tsunami warning system: A case study in Indonesia's natural disaster management*.

Chen, E., Lerman, K., & Ferrara, E. (2020a). Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*.

Chen, E., Lerman, K., & Ferrara, E. (2020b). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance, 6*(2), e19273.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information storage and retrieval, 7*(1), 19-37.

Das, S. (1994). *Time series analysis* (Vol. 10): Princeton university press, Princeton, NJ.

Djalante, R., Shaw, R., & DeWit, A. (2020). Building resilience against biological hazards and pandemics: COVID-19 and its implications for the Sendai Framework. *Progress in Disaster Science*, 100080.

Dünser, M. W., Baelani, I., & Ganbold, L. (2006). A review and analysis of intensive care medicine in the least developed countries. *Critical care medicine, 34*(4), 1234-1242.

Elliott, C. (1998). Defining development news values: An examination of press releases from the New China News Agency. *Mass media in the Asian Pacific*, 72-84.

Fauci, A. S., Lane, H. C., & Redfield, R. R. (2020). Covid-19—navigating the uncharted. In: Mass Medical Soc.

Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications, 68*(13), 13-18.

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., . . . Hui, D. S. C. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England journal of medicine, 382*(18), 1708-1720.

Gurman, T. A., & Ellenberger, N. (2015). Reaching the global community during disasters: findings from a content analysis of the organizational use of Twitter after the 2010 Haiti earthquake. *Journal of health communication, 20*(6), 687-696.

Haynes, W. (2013). Student's t-test. *Encyclopedia of Systems Biology*, 2023-2025.

Hsu, H., & Lachenbruch, P. A. (2005). Paired t test. *Encyclopedia of Biostatistics, 6*.

IFLA. (2017). List of Countries in Asia & Oceania. Retrieved from https://www.ifla.org/node/9511

IMF. (2000). UN recognition of the least developed countries. Retrieved from https://unctad.org/topic/least-developed-countries/recognition

Ishiwatari, M., Koike, T., Hiroki, K., Toda, T., & Katsube, T. (2020). Managing disasters amid COVID-19 pandemic: Approaches of response to flood disasters. *Progress in Disaster Science*, 100096.

Islam, S. M. D.-U., Bodrud-Doza, M., Khan, R. M., Haque, M. A., & Mamun, M. A. (2020). Exploring COVID-19 stress and its factors in Bangladesh: A perception-based study. *Heliyon, 6*(7), e04399.

Jackendoff, R. (1992). *Semantic structures* (Vol. 18): MIT press.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). *Why we twitter: understanding microblogging usage and communities*.

Kalra, V., & Aggarwal, R. (2017). *Importance of Text Data Preprocessing & Implementation in RapidMiner*.

Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16.

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*: Springer Science & Business Media.

Kasper, W., & Vela, M. (2011). *Sentiment analysis for hotel reviews*.

Kireyev, K., Palen, L., & Anderson, K. (2009). *Applications of topics models to analysis of disaster-related twitter data*.

Kongthon, A., Haruechaiyasak, C., Pailai, J., & Kongyoung, S. (2014). The role of social media during a natural disaster: A case study of the 2011 Thai Flood. *International Journal of Innovation and Technology Management, 11*(03), 1440012.

Kraemer, M. U. G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., . . . Hanage, W. P. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science, 368*(6490), 493-497.

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). *Cosine similarity to determine similarity measure: Study case in online essay assessment*.

Lantz, P. M., & Booth, K. M. (1998). The social construction of the breast cancer epidemic. *Social science & medicine, 46*(7), 907-918.

Leelawat, N., Tang, J., Saengtabtim, K., & Laosunthara, A. (2020). Trends of Tweets on the Coronavirus Disease-2019 (COVID-19) Pandemic. *Journal of Disaster Research, 15*(4), 530-533.

Leung, C. C., Lam, T. H., & Cheng, K. K. (2020). Mass masking in the COVID-19 epidemic: people need guidance. *Lancet, 395*(10228), 945.

Li, B., & Han, L. (2013). *Distance weighted cosine similarity measure for text classification*.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., . . . Wong, J. Y. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England journal of medicine*.

Li, S., Wang, Y., Xue, J., Zhao, N., & Zhu, T. (2020). The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. *International journal of environmental research and public health, 17*(6), 2032.

Li, W., Feng, Y., Li, D., & Yu, Z. (2016). Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm. *Automatic Control and Computer Sciences, 50*(4), 271-277.

Lim, S. S., Allen, K., Bhutta, Z. A., Dandona, L., Forouzanfar, M. H., Fullman, N., . . . Holmberg, M. (2016). Measuring the health-related Sustainable Development Goals in 188 countries: a baseline analysis from the Global Burden of Disease Study 2015. *The Lancet, 388*(10053), 1813-1850.

Lin, J., & Wilbur, W. J. (2007). PubMed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics, 8*(1), 423.

Liu, N., Zhang, B., Yan, J., Yang, Q., Yan, S., Chen, Z., . . . Ma, W.-Y. (2004). *Learning similarity measures in non-orthogonal space*.

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics, 11*(1-2), 22-31.

MacMullen, W. J. (2003). Requirements definition and design criteria for test corpora in information science. *on-line all" indirizzo:* http://sils*. unc. edu/sites/default/files/general/research/TR-2003-03. pdf (visitato 02/01/2010)*.

Makice, K. (2009). *Twitter API: Up and running: Learn how to build applications with the Twitter API*: " O'Reilly Media, Inc.".

Mathioudakis, M., & Koudas, N. (2010). *Twittermonitor: trend detection over the twitter stream*.

Meng, J., & Berger, B. K. (2008). Comprehensive dimensions of government intervention in crisis management: A qualitative content analysis of news coverage of the 2003 SARS epidemic in China. *China Media Research, 4*(1), 19-28.

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*.

Murray, C. J. L. (2015). Choosing indicators for the health-related SDG targets. *The Lancet, 386*(10001), 1314-1317.

Novel, C. P. E. R. E. (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. *Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi, 41*(2), 145.

Novoselov, S., Shchemelinin, V., Shulipa, A., Kozlov, A., & Kremnev, I. (2018). *Triplet Loss Based Cosine Similarity Metric Learning for Text-independent Speaker Recognition*.

of the International, C. S. G. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology, 5*(4), 536.

Palliyaguru, R., Amaratunga, D., & Baldry, D. (2014). Constructing a holistic approach to disaster risk reduction: the significance of focusing on vulnerability reduction. *Disasters, 38*(1), 45-61.

Perlman, S. (2020). Another decade, another coronavirus. In: Mass Medical Soc.

Peters, D. H., Hanssen, O., Gutierrez, J., Abrahams, J., & Nyenswah, T. (2019). Financing common goods for health: Core government functions in health emergency and disaster risk management. *Health Systems & Reform, 5*(4), 307-321.

Poissy, J., Goffard, A., Parmentier-Decrucq, E., Favory, R., Kauv, M., Kipnis, E., . . . The, M. (2014). Kinetics and pattern of viral excretion in biological specimens of two MERS-CoV cases. *Journal of Clinical Virology, 61*(2), 275-278.

Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., & Mohanty, S. (2019). Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. *International journal of disaster risk reduction, 37*, 101176.

Quarantelli, E. L. (1988). Disaster crisis management: A summary of research findings. *Journal of management studies, 25*(4), 373-385.

Raghuveer, K. (2012). Legal documents clustering using latent dirichlet allocation. *IAES Int. J. Artif. Intell, 2*(1), 34-37.

Ramos, J. (2003). *Using tf-idf to determine word relevance in document queries*.

Ramya, R. S., Sejal, D., Venugopal, K. R., Iyengar, S. S., & Patnaik, L. M. (2018). *Drdlc: Discovering Relevant Documents using Latent Dirichlet Allocation and Cosine Similarity*.

Roesslein, J. (2015). Tweepy. *Python programming language module*.

Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020). Coronavirus pandemic (COVID-19). *Our World in Data*.

Rufai, S. R., & Bunce, C. (2020). World leaders’ usage of Twitter in response to the COVID-19 pandemic: a content analysis. *Journal of Public Health, 42*(3), 510-516.

Rus, V., Niraula, N., & Banjade, R. (2013). *Similarity measures based on latent dirichlet allocation*.

Rusu, O., Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M., & Marinescu, V. (2013). *Converting unstructured and semi-structured data into knowledge*.

Ryu, W.-S. (2016). *Molecular virology of human pathogenic viruses*: Academic Press.

Sakai, T. (2016). *Two sample t-tests for ir evaluation: Student or welch?*

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors*.

Scott, K. K., & Errett, N. A. (2018). Content, accessibility, and dissemination of disaster information via social media during the 2016 Louisiana floods. *Journal of public health management and practice, 24*(4), 370-379.

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media, 22*, 100104.

Simpson, D. M., & Usey Jr, R. W. (2004). System and method for building a semantic network capable of identifying word patterns in text. In: Google Patents.

Subasi, A., Kevric, J., & Canbaz, M. A. (2019). Epileptic seizure detection using hybrid machine learning methods. *Neural Computing and Applications, 31*(1), 317-325.

The Sustainable Development Goals and Addressing Statelessness. (March 2017). Retrieved from

Tan, A.-H. (1999). *Text mining: The state of the art and the challenges*.

Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM Sigmod Record, 36*(2), 7-12.

Tirkkonen, P., & Luoma-aho, V. (2011). Online authority communication during an epidemic: A Finnish example. *Public Relations Review, 37*(2), 172-174.

UNCTAD. (2020). UN recognition of the least developed countries. Retrieved from https://unctad.org/topic/least-developed-countries/recognition

Unisdr, U. (2015). *Sendai framework for disaster risk reduction 2015–2030*.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., . . . Fu, H. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*.

Verma, T., Renu, R., & Gaur, D. (2014). Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems, 7*(2), 16-18.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16.

Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal, 3*(2), 19-28.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). *A system for real-time twitter sentiment analysis of 2012 us presidential election cycle*.

Wang, L.-s., Wang, Y.-r., Ye, D.-w., & Liu, Q.-q. (2020). A review of the 2019 Novel Coronavirus (COVID-19) based on current evidence. *International journal of antimicrobial agents*, 105948.

WHO. (2020). *Novel Coronavirus (2019-nCoV) SITUATION REPORT - 1*. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4

World Health, O. (2015). Health in 2015: from MDGs, millennium development goals to SDGs, sustainable development goals.

World Health, O. (2016). *World health statistics 2016: monitoring health for the SDGs sustainable development goals*: World Health Organization.

World Health, O. (2019). Health emergency and disaster risk management framework.

World Health, O. (2020). *Situation report, 22*. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf

Worldometers. (24 September, 2020). Retrieved from https://www.worldometers.info/

Wright, N., Fagan, L., Lapitan, J. M., Kayano, R., Abrahams, J., Huda, Q., & Murray, V. (2020). Health emergency and disaster risk management: Five years into implementation of the Sendai Framework. *International Journal of Disaster Risk Science, 11*(2), 206.

Ye, J. (2011). Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Mathematical and computer modelling, 53*(1-2), 91-97.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied artificial intelligence, 17*(5-6), 375-381.

Zhang, Y., Lyu, H., Liu, Y., Zhang, X., Wang, Y., & Luo, J. (2020). Monitoring Depression Trend on Twitter during the COVID-19 Pandemic. *arXiv preprint arXiv:2007.00228*.

VITA

NAME	Kumpol Saengtabtim
DATE OF BIRTH	05 December 1996
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	SIIT Thammasart University (B.Sc) Chulalongorn University
HOME ADDRESS	113 Charansanitwong road , Charansanitwong 75, Soi 6 , Bang plat district, Bang O , Bangkok, 10700
PUBLICATION	Trends of Tweets on the Coronavirus Disease-2019 (COVID-19) Pandemic., Predictive Analysis of the Building Damage From the 2011 Great East Japan Tsunami Using Decision Tree Classification Related Algorithms., Today in Thailand: multidisciplinary perspectives on the current tsunami disaster risk reduction., Twitter Sentiment Analysis of Bangkok Tourism During COVID-19 Pandemic Using Support Vector Machine Algorithm., Effectiveness of Applying HEALTH-EDRM Framework: A Comparison of the COVID-19 Situation in Asia-Oceania Countries and Territories.
AWARD RECEIVED	Bhumibol Scholarship(Gold medal in Engineering management).