1. Introduction

1.1 Background

The term infodemiology was invented by Gunther Eysenbach in 2002 and described as "a new area of scientific research that studies the distribution and determinants of information in an electronic medium, especially the internet, for user-contributed health-related content with the ultimate goal of improving public health" (Eysenbach, 2002). In 2004, the term infoveillance appeared to distinguish the type of study and the activity (Eysenbach, 2009). While infodemiology was perceived as the study, infoveillance was regarded as the surveillance activity that was conducted to learn infodemiology, whether past events or in real-time (Colditz *et al*., 2018) (Abd-Alrazaq *et al*., 2020). Infodemiology is also synonymous with digital epidemiology or digital disease detection (DDD) (O'Shea, 2017). Nowadays, infodemiology is defined as a new discipline in big data that utilizes data generated outside the public health system to understand the patterns and determinants of health dynamics in populations to mitigate and prevent disease alongside promoting health (Salathé, 2018). It promises faster detection of outbreaks and improved surveillance along with reductions in administrative and financial burden by utilizing data that wasn't primarily generated to do epidemiology; examples include a mobile phone, search engines, website access logs, and social media services (Eckmanns *et al.*, 2019). Google Flu Trends (GFT) was the earliest known example of infodemiology: it worked by utilizing symptomatic search queries for tracking influenza-like illness (Ginsberg *et al*., 2009). The limitation of using symptomatic search queries and google trends, such as GFT, was already addressed in a prior study (Cook *et al.,* 2011; Olson *et al.*, 2013; King *et al.*, 2014; and Cervellin *et al*., 2017). The latter innovative approaches coming out were the establishment of disease monitoring by academics labs such as HealthMap (Brownstein *et al*., 2008), InfluenzaNet (Paolotti *et al*., 2014), and the Coronavirus COVID-19 Global Cases mapping by Johns Hopkins CSSE (Zlojutro *et al*., 2019) (Dong *et al*., 2020). In addition to data-related trends, worldwide drug use analysis by utilizing google trends has also been applied (Lippi *et al.,* 2017). Beyond the data-related trends, utilization of machine learning through its subfields called deep learning (LeCun *et al*., 2015) builds on artificial neural networks to learn arduous amounts of data to map the input to output correctly, has entered the world of infodemiology. The example was: a study on almost 130,000 clinical images of skin lesions performed on par with 21 board-certified dermatologists to classify skin cancer (Esteva *et al.,* 2017) and a deep learning system being compared to dermatologists, primary care physicians, and nurse practitioners to detect 26 skin conditions based from 7 years data training (Liu *et al*., 2019).

Despite its numerous initiatives, infodemiology has its most significant limitation from the private ownership of the data source. It is aligned with the current trend of primary internet services, which substantially reducing access to their data. As an example, Instagram since 2012 has severely restricted access within their application programming interface (API) (Salathé, 2018). Will Twitter remain openly accessible is unbeknown, but still, the decision remains in the hands of the owners of Twitter Inc.

Infodemiology, perhaps, will become an essential part of future disease prevention, health-related risk mitigation, and an adjacent form of the health-promoting platform, waiting to be matured and fully integrated into the daily workflow of public health authorities. In adjoining to these beliefs, a prior study has been conducted to explore diffusion and the trends of infodemiology by carrying Spearman correlation analysis based on Medline from 2012 until 2018 (Lippi & Cervellin, 2019). The results show that there is an ongoing shift within clinical epidemiology. In conclusion, infodemiology cannot replace clinical epidemiology. However, the innovative discipline shall be regarded as proper support, which can be utilized and optimized to generate timely alerts for clinical epidemiologists on disease outbreaks sudden changes of therapeutic options, much earlier than conventional health epidemiology.

The 2019 novel coronavirus disease (COVID-19), informally referred to as Wuhan coronavirus or 2019-nCoV for its virus, belongs to a family of single-stranded RNA viruses known as coronaviridae (Zhou *et al.,*2020). COVID-19 infection caused clusters of severe respiratory illness with major symptoms: including fever, cough, and myalgia or fatigue and minor symptoms: including sputum production, headache, hemoptysis, and diarrhea with the result of dyspnoea, lymphopenia, and pneumonia (Huang *et al*., 2020). The United States Centre for Disease Control has indicated that the symptoms appear in a few around two until 14 days after the exposure (Centre for Disease Control, 2020). However, COVID-19 can infect others within its incubation period (Siddique *et al*., 2020). On January 23, 2020, the World Health Organization (WHO) has estimated that the death rate of COVID-19 will be at 3%; however, the preliminary R0 (reproduction number) is being estimated at 1.4 to 2.5, meaning that every person infected could infect between 1.4 to 2.5 people (Mahase, 2020). Like other common respiratory tract infections, COVID-19 was spread through respiratory droplets produced by an infected person when they sneeze or cough with the possibilities of fecal as other means of disease transmission (Chan *et al*., 2020) (Pinghui, 2020). Due to its character of low mortality but easily transmitted, within March 6, pass two month after it initial report on December 31, 2019, COVID-19 has affected 101,800 person with over 100 countries, resulting in 3460 deaths with 55,866 patients able to recover (Dong *et al*., 2020). COVID-19 rapid transmission, along with its status as a new strain of the virus, has led a rush within academia, business media, community society, and government official to understand and share the updates of the virus itself. This rush led humanity to battle epidemiology and misinformation and hoax regarding the disease, as stated by Tedros Adhanom Ghebreyesus, WHO Director-General, at the Munich Security Conference on February 15. “We’re not just fighting an epidemic; we’re fighting an infodemic!” (WHO, 2020a). In a respond to *The Lancet*, Sylvie Briand, director of Infectious Hazards Management at WHO’s Health Emergencies Programme and architect of WHO strategy to counter infodemic risk; stated infodemic as misinformation, rumors, etc. that were part of a tsunami of information following every outbreak, even the record of it can be tracked back into Middle Ages Era (Zarocostas, 2020). In combating infodemics, WHO itself has managed to contact and cooperate with various social media such as Twitter, Facebook, Pinterest, and Wechat to direct official information myth-buster links from WHO and Centers for Disease Control and Prevention when people browse information related to corona. For example, Facebook has removed a modified WHO infographic that claims people should avoid having sex with animals to prevent coronavirus (Ritchel, 2020).

During the COVID-19 pandemic, the term infodemiology has been re-defined by Stefano Burzo at the 1st WHO infodemiology conference as "the science of managing infodemics." (WHO, 2020c) Infodemic itself was defined as an overabundance of information, whether accurate or not, making it harder for people to find trustworthy sources and reliable guidance when needed. WHO had considered infodemiology as a multi-disciplinary subject, with the aspects were: epidemiology and public health; media studies and journalism; risk communication and community engagement; ethics, law, and governance; applied math and data science; digital health and technology applications; social behavior and science; marketing; user experience and design (Tangcharoensathien, 2020).

1.2 Objective

In this study, the author conducted an infoveillance study towards COVID-19 infodemiology distribution on Indonesia, Singapore, and the United States on its government and news media youtube channel to study the differences of behavior in information response between the three countries and the outcomes of it. The study utilizes text analytics methods based on natural language understanding, including topic detection, topic classification, and sentimental analysis, with timeline information adjacent to COVID-19 regulation and case update. Our data consisted of youtube content published from January - March 2020 with the government official youtube channels, three news media of each country, three international news media, and WHO as the study's subject.

This study aims to conduct an infoveillance study towards Indonesia, Singapore, and the United States responds towards the COVID-19 pandemic using text analytics methods based on circulating data on Youtube. The reason youtube was chosen was due to its scarcity as being the source for an infodemiology study. On the other hand, it represents the content of local-air television that being accessed as the most consumed media with an increase of over 50% in Indonesia, 133% in Singapore, and 73% in the United States during the pandemic (Lukman, 2020) (GFK, 2020) (Sullivan & Molay, 2020). This study focuses on the government and news media's information response to compare the response, sentiment over time, and content coverage, which have been uploaded by different sources during the early phase of the pandemic.

In previous studies, assessment of information credibility of COVID-19 was focused only on news media in a specific country as towards news media in China and Korea (Liu *et al.,* 2020) (Kim, 2020). Coverage towards government publication was scarce, and text analytics conducted rarely merge sentiment analysis with emotional analysis (Aslam *et al*., 2020). This study address 4 out of 5 takeaways from Tim Zecchin, Managing Director of Media Measurement Ltd. (https://www.mediameasurement.com/), presentation at WHO 1st Infodemiology Conference: Analyzing emotion and sentiment, identifying emerging narratives, the volume of conversation and velocity detection, and measuring the level of correct information; with the point of takeaways that were not able to detect was virality of the content along with the emerging topics that came from citizen as feedback. As a result, this study delivers significances on the utilization of various text-analytics methods that previously were segregated and the detection of infodemiology towards multiple countries rather than focusing only on a single one and using youtube as the dataset that was scarce compared to social media and news articles.

1.3 Hypothesis

The author hypothesized that amongst the three-nation, Singapore would result in the best response and overall quality compared to another country due to its better political stability and since its healthcare system has been regarded as the most efficient (Bloomberg, 2014) (Ortmann & Thompson, 2018). Content coverage will vary among countries with more coherent but with similar sentiment and narrative over the pandemic's timeline progress.

2. Literature Review

2.1 Infodemiology study based on text analysis

Infodemiology study based on text analytics has been conducted through various sources of data and variant purposes. Within news media and the scientific publication, a prior study has compared the contents of both publications on ebola topics by using text-mining techniques. It resulted in topic modeling, word co-occurrence map, and network analysis, showing differences and commonalities between social coverage and academic research (An *et al.,* 2016). Regarding the government or public health official statements, an analysis of past and present US presidents' words and actions has confirmed Donal Trump's hostility towards science. It suggests it may hurt him politically as an outlier (Quiley *et al*., 2020). A measurement of the frequency of 30 different keywords and their derivatives on president speech transcript and presidential budget messages as well the positive statistical relationship between science advocacy and political popularity among US Presidents with cluster tree establishment has indicated how Trump has abdicated science through his official speech across his reign.

Regarding public response, Twitter has undoubtedly become the most explored infodemiology source with its coverage of over 275 million users, and around 330 million of them were active users (Broniatowski *et al*., 2014; Clement, 2019a; Clement 2019b). In a prior study, Twitter has been used to track rapidly-evolving public sentiment concerning H1N1 or swine flu, track and measure actual disease activity, and showcase how it can be accurately used to estimate influenza-like illness within the United States of America (Signorini *et al*., 2011). Due to its robustness, Twitter infodemiology study has become a common source with 4117 publications listed on Pubmed as of September 2020. Less famous infoveillance platform sources resulted in lower numbers such as 3951 for Facebook, 1473 for Youtube, and 579 for Instagram.