1-2022

Exploring how online responses change in response to debunking messages about COVID-19 on WhatsApp

Xingyu Ken CHEN
Nanyang Technological University

Jin-Cheon NA
Nanyang Technological University

Luke Kien-Weng TAN
Nanyang Technological University

Mark CHONG
Singapore Management University, markchong@smu.edu.sg

Murphy CHOY
National University of Singapore

Citation
CHEN, Xingyu Ken; NA, Jin-Cheon; TAN, Luke Kien-Weng; CHONG, Mark; and CHOY, Murphy. Exploring how online responses change in response to debunking messages about COVID-19 on WhatsApp. (2022). *Online Information Review*. 46, (6), 1184-1204.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/6973

Exploring how online responses change in response to debunking messages about COVID-19 on WhatsApp

Xingyu Ken Chen (Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore)

Jin-Cheon Na (Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore)

Luke Kien-Weng Tan (Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore)

Mark Chong (Lee Kong Chian School of Business, Singapore Management University, Singapore, Singapore)

Murphy Choy (Alionova Consulting Private Limited, Singapore, Singapore)

Published in Online Information Review, 2022, DOI: 10.1108/OIR-08-2021-0422

Abstract

Purpose

The COVID-19 pandemic has spurred a concurrent outbreak of false information online. Debunking false information about a health crisis is critical as misinformation can trigger protests or panic, which necessitates a better understanding of it. This exploratory study examined the effects of debunking messages on a COVID-19-related public chat on WhatsApp in Singapore.

Design/methodology/approach

To understand the effects of debunking messages about COVID-19 on WhatsApp conversations, the following was studied. The relationship between source credibility (i.e. characteristics of a communicator that affect the receiver's acceptance of the message) of different debunking message types and their effects on the length of the conversation, sentiments towards various aspects of a crisis, and the information distortions in a message thread were studied. Deep learning techniques, knowledge graphs (KG), and content analyses were used to perform aspect-based sentiment analysis (ABSA) of the messages and measure information distortion.

Findings

Debunking messages with higher source credibility (e.g. providing evidence from authoritative sources like health authorities) help close a discussion thread earlier. Shifts in sentiments towards some aspects of the crisis highlight the value of ABSA in monitoring the effectiveness of debunking messages. Finally, debunking messages with lower source credibility (e.g. stating that the information is false without any substantiation) are likely to increase information distortion in conversation threads.

Originality/value

The study supports the importance of source credibility in debunking and an ABSA approach in analysing the effect of debunking messages during a health crisis, which have practical value for public agencies during a health crisis. Studying differences in the source credibility of debunking messages on WhatsApp is a novel shift from the existing approaches. Additionally, a novel approach to measuring information distortion using KGs was used to shed insights on how debunking can reduce information distortions.

Keywords: COVID-19, Debunking, Aspect-based sentiment analysis, Information distortion, Source credibility, Deep learning

Introduction

The outbreak of COVID-19 is an ongoing pandemic first reported in Wuhan, China (Origin of SARS-CoV-2, 2020). Misinformation also spread concurrently with the outbreak on a global scale, causing profound impacts on society. Notably, claims about 5G technologies causing the spread of COVID-19 triggered protests in Australia and the United Kingdom (Meese *et al.*, 2020). Singapore is not immune to the spread of false information either. For instance, an article by an alternative media outlet falsely claimed that face masks had run out (Chong, 2020), which could create public panic. Moreover, COVID-19-related misinformation also contributed to increased xenophobic and racist sentiments (Abdul Rahman, 2020).

Consequently, the COVID-19 infodemic attracted considerable scholarly interest in various research areas, such as uncovering motivations for sharing fake news on WhatsApp (Apuke and Omar, 2020) or how COVID-19 misinformation manifests in different countries (Zeng and Chan, 2021).

It is vital to the public interest to counter false information during a health crisis. Consequently, debunking efforts have grown in response to the spread of false information. For instance, over 100 fact-checking organisations came together to create a database of fact-checked information for COVID-19 ("CoronaVirusFacts Alliance", no date). As an area of study, debunking false information is attracting scholarly interest to address questions like how it can reduce belief in misinformation (Walter *et al.*, 2020) or reduce rumour spread (Ermakova *et al.*, 2020; Jiang *et al.*, 2021). In fact, recent systematic reviews found that corrections could significantly reduce belief in misinformation compared to cases where the misinformation was not debunked (Chan *et al.*, 2017; Walter *et al.*, 2020).

Despite the plethora of COVID-19 research concerning the infodemic, how do debunking messages about COVID-19 affect WhatsApp conversations is currently understudied. This question is important because, according to some studies, these messaging apps can exert much more social influence than social media sites (Rouhani *et al.*, 2019; Yu and Poger, 2019). COVID-19 conversations on WhatsApp are understudied compared to studies on social media platforms like Twitter (see Ermakova *et al.*, 2020; Jiang *et al.*, 2021), as COVID-19 chats are rarer due to their closed nature. Observational studies on the role of debunking in natural COVID-19-related WhatsApp conversations are also scarce. Other studies frequently use experimental/survey research designs to study this (Bowles *et al.*, 2020; Vijaykumar *et al.*, 2021) rather than using observational designs.

For debunking, there are also other variables of interest. For practitioners, source credibility is essential in debunking. For COVID-19, the World Health Organisation advised the need for community engagement rooted in evidence-based, open communications from trusted sources (COVID-19 global risk communication and community engagement strategy, December 2020-May 2021: interim guidance, 2020). In Singapore, the authorities used various social media platforms, including WhatsApp and the official Health Ministry website, to disseminate COVID-19 updates to inform the population and counteract misinformation (Basu, 2020).

Furthermore, few studies discuss how debunking influences sentiments (Zeng and Zhu, 2019; Jang *et al.*, 2021) or its effects on variables like the length of the conversation or the amount of information distortion. As these are not studied together, we used innovative approaches to study how debunking affects WhatsApp conversations on the COVID-19 pandemic in Singapore and the variables above.

Related work

Information disorder and COVID-19

"Fake news" has been used as a blanket term to describe misinformation, disinformation and malicious information across various settings, such as climate change and vaccine hesitancy. However, the term is vulnerable to misuse by political figures, highlighting a need to use "information disorder" (IDO) as a replacement (Wardle and Derakhshan, 2017). Hence, this study used IDO to mean content containing or referencing an instance of disinformation/misinformation/mal-information as described by Wardle and Derakhshan (2017).

Substantial literature is devoted to detecting IDO (see review by de Souza *et al.*, 2020), spurred by successes in deep learning techniques applied to detecting IDO. One example is bidirectional encoder representations from transformers (BERT). BERT is a bidirectional general language model pre-trained on large-scale unsupervised text corpus (Devlin *et al.*, 2018), which can be fine-tuned for text classification. For COVID-19, Ayoub *et al.* (2021) developed an explainable natural language processing model based on DistilBERT and SHAP to detect COVID-19 misinformation with good performance on various COVID-19 fake news data sets where the accuracy is above 0.9. DistilBERT is part of a family of BERT-based models which have strong performance on fake news detection tasks (see Aggarwal *et al.*, 2020; Baruah *et al.*, 2020). Thus, this study used BERT-based models for text classification.

Debunking and source credibility

Source credibility is critical for countering misinformation, as relevant research in health communication of risks found that reliable sources help in improving belief accuracy (van der Meer and Jin, 2020; Vijaykumar *et al.*, 2021). However, current research on the role of source credibility in moderating the effects of a debunking message on belief in misinformation is mixed (Walter and Tukachinsky, 2019; Vijaykumar *et al.*, 2021). According to Lewandowsky *et al.* (2021), people do not routinely track source credibility. However, when they do, the effect of IDO from less credible sources can be reduced.

Moreover, the persuasiveness of debunking messages can vary based on source credibility (see Bordia *et al.*, 2005; Pornpitakpan, 2004). Thus, this study proposes to measure how differences in the source credibility of debunking messages on WhatsApp conversations influence variables such as the conversation thread length, sentiment and information distortion. Examining source credibility and these variables in tandem can yield novel insights for the literature.

Aspect-based sentiment analysis

Crises that threaten the public's safety are emotionally trying even if it is experienced vicariously (Coombs, 2015; Chong and Choy, 2018). Sentiment analysis (SA) has been used to assess the public's sentiments and emotions during crises, including COVID-19 (Alamoodi *et al.*, 2021). Singh *et al.* (2021) used BERT to analyse tweets in India and associated positive sentiments with events such as government control of the virus and negative sentiments towards lockdown measures. Furthermore, there is evidence of the link between emotions and belief in false information. Those who use emotions to assess a news story are more likely to believe false news (Martel *et al.*, 2020).

Aspect-based sentiment analysis (ABSA) can yield more granular and practical analysis as it identifies sentiment polarity towards an aspect category rather than the entire message or sentence. Many studies have used ABSA (see review by Do *et al.*, 2019). In particular, Jang *et al.* (2021) used this to derive valuable insights towards aspects of COVID-19. They found that Twitter users' reactions towards aspects such as misinformation appeared to be more negative than positive, suggesting frustration with misinformation. Using pre-trained, attention-based neural models to examine words surrounding target aspects have yielded promising results. For instance, Sun *et al.* (2019) used BERT for ABSA, achieving a new state-of-the-art on SentiHood and SemEval-2014 Task 4 data set. Hence, this study will also leverage BERT for carrying out ABSA-related tasks.

Information distortion

A message often undergoes changes as people spread, dissect and interpret it. Analogous to the telephone game, content gets increasingly distorted as it spreads. Thus, for researchers studying information distortion (IDT), quantifying the integrity of information as it diffuses through a medium can yield fruitful analysis.

Using experimental diffusion chains, Moussaïd *et al.* (2015) showed how messages about the risks of a controversial antibacterial agent change when transmitted through a ten-person diffusion chain. Another approach involves tracking changes to selected words along an information cascade. In the COVID-19 setting, Ermakova *et al.* (2020) tracked medical IDT occurring in cascades on Twitter by identifying substitutions for selected words as a way to detect IDT. Nevertheless, the aforementioned methods of measuring IDT cannot be transplanted to the WhatsApp context. Group conversations are rarely iterative summarisations of previous messages, and group members could contribute new information or go off-topic before continuing the previous discussion. Hence, this study proposes a new method for identifying IDTs by combining knowledge graphs, numerical relation extraction, and IDO labelling.

Research objectives

In order to understand how debunking messages about COVID-19 in Singapore affected WhatsApp conversations, we aim to address the current research gaps through the following research questions:
RQ1. What effects do differences in the source credibility of debunking messages have on WhatsApp conversations? Identifying source credibility differences in debunking messages widens the current evidence base for debunking. Additionally, understanding how debunking affects the dynamics of a WhatsApp conversation, such as reducing an information cascade, influencing differences in sentiment or reducing IDT, contributes to the evidence for debunking-related interventions.
RQ2. How do sentiments change in response to debunking messages? Using an aspect-based SA provides a fine-grained analysis of how debunking messages influences sentiments at the aspect level. Additionally, showing how debunking messages influence sentiments towards various aspects of a COVID-19 crisis is a novel contribution to existing research for debunking and SA.
RQ3. How does IDT change in response to debunking messages? As this is an under-studied area, measuring IDT on conversation threads can provide novel insights and contributions to the literature.

Methods

To address the three RQs, a combination of deep learning techniques, knowledge graphs (KG), and content analysis were used (see Figure 1). Three human coders (A, B, C) were used for all manual content analysis methods. They resided in Singapore from Feb 2020–Apr 2020 and are familiar with the events discussed in the dataset.

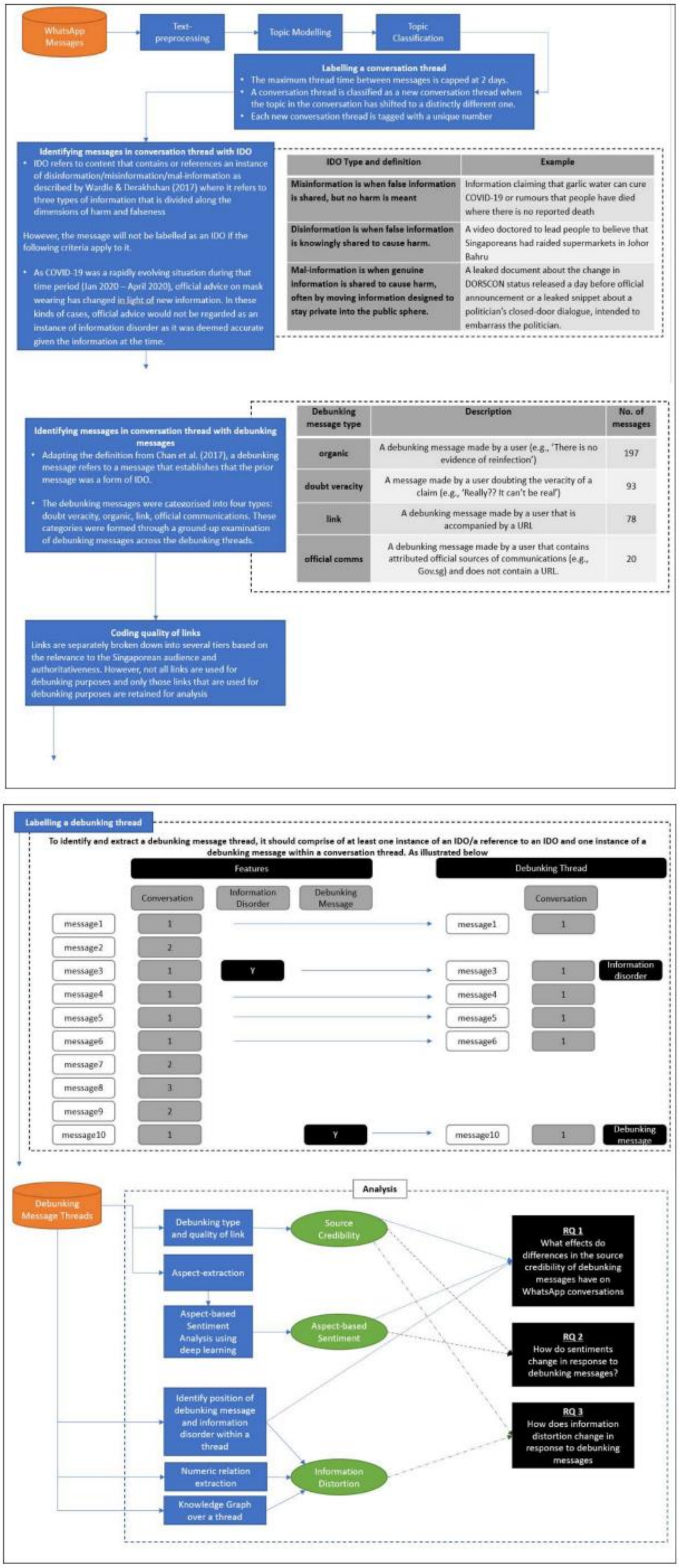
Data collection

The dataset was from discussions about COVID-19 in Singapore captured in a public WhatsApp group administered as part of a public forum. There were 50,187 messages and 209 users spanning Feb 2020 to Apr 2020. This period marked the initial phases of the coronavirus outbreak in Singapore, where the Singapore Government began rolling out measures to combat COVID-19, such as the restriction of travellers, cancellation of events, raising of the DORSCON (Disease Outbreak Response System Condition) status to DORSCON Orange and implementation of Circuit Breaker Measures. During this period, there was debate over the effectiveness of mask-wearing. One other major external event that significantly affected Singapore was Malaysia's announcement of a Movement Control Order (MCO).

To protect the research subject's privacy, this study used methods that were also used in other studies. Firstly, similar studies (see Bursztyn and Birnbaum, 2019; Garimella and Tyson, 2018) highlight that according to the current WhatsApp privacy policy, users should be aware that when posting a message on a group chat, that message is accessible to other members of the group, which is particularly true for messages on a publicly available group chat which allows people. Secondly, users who participate in this group chat are aware that such discussions are collected for feedback based on the group's terms of use which are regularly posted in the chat. Thirdly, identifying information was anonymised before any analysis (e.g. phone numbers of users). No content posted can be linked to any personal data, and thus personal data cannot inform research findings. This is consistent with recommendations by Barbosa and Milan (2019) on the ethics surrounding research on WhatsApp.

Pre-processing the text

Text normalisation was performed using a dictionary-based approach, and non-alphanumeric characters were removed. Singlish phrases (e.g. “bo pian”[“no choice”]) were translated to English. In addition, various acronyms were expanded and underscored to reduce the dimensionality of the data set. Each message was broken down into sentences using NLTK's sentence tokenizer for more fine-grained analysis, rather than using the entire message as a unit of analysis for topic classification (TC) and SA. This approach is common in SA, where document-level or sentence-level analyses are conducted (Do *et al.*, 2019).



Labelling a conversation thread

- The maximum thread time between messages is capped at 2 days.
- A conversation thread is classified as a new conversation thread when the topic in the conversation has shifted to a distinctly different one.
- Each new conversation thread is tagged with a unique number

Identifying messages in conversation thread with IDO

- IDO refers to content that contains or references an instance of disinformation/misinformation/mal-information as described by Wardle & Derakhshan (2017) where it refers to three types of information that is divided along the dimensions of harm and falseness

However, the message will not be labelled as an IDO if the following criteria apply to it.

- As COVID-19 was a rapidly evolving situation during that time period (Jan 2020 – April 2020), official advice on mask wearing has changed in light of new information. In these kinds of cases, official advice would not be regarded as an instance of information disorder as it was deemed accurate given the information at the time.

IDO Type and definition	Example
Misinformation is when false information is shared, but no harm is meant	Information claiming that garlic water can cure COVID-19 or rumours that people have died where there is no reported death
Disinformation is when false information is knowingly shared to cause harm.	A video doctored to lead people to believe that Singaporeans had raided supermarkets in Johor Bahru
Mal-information is when genuine information is shared to cause harm, often by moving information designed to stay private into the public sphere.	A leaked document about the change in DORSCON status released a day before official announcement or a leaked snippet about a politician's closed-door dialogue, intended to embarrass the politician.

Identifying messages in conversation thread with debunking messages

- Adapting the definition from Chan et al. (2017), a debunking message refers to a message that establishes that the prior message was a form of IDO.
- The debunking messages were categorised into four types: doubt veracity, organic, link, official communications. These categories were formed through a ground-up examination of debunking messages across the debunking threads.

Debunking message type	Description	No. of messages
organic	A debunking message made by a user (e.g., 'There is no evidence of reinfection')	197
doubt veracity	A message made by a user doubting the veracity of a claim (e.g., 'Really?? It can't be real')	93
link	A debunking message made by a user that is accompanied by a URL	78
official comms	A debunking message made by a user that contains attributed official sources of communications (e.g., Gov.sg) and does not contain a URL.	20

Coding quality of links

Links are separately broken down into several tiers based on the relevance to the Singaporean audience and authoritativeness. However, not all links are used for debunking purposes and only those links that are used for debunking purposes are retained for analysis

Labelling a debunking thread

To identify and extract a debunking message thread, it should comprise of at least one instance of an IDO/a reference to an IDO and one instance of a debunking message within a conversation thread. As illustrated below

	Features			Debunking Thread		
	Conversation	Information Disorder	Debunking Message		Conversation	
message1	1			message1	1	
message2	2					
message3	1	Y		message3	1	Information disorder
message4	1			message4	1	
message5	1			message5	1	
message6	1			message6	1	
message7	2					
message8	3					
message9	2					
message10	1		Y	message10	1	Debunking message

Debunking Message Threads
Analysis
Debunking type and quality of link
Aspect-extraction
Aspect-based Sentiment Analysis using deep learning
Identify position of debunking message and information disorder within a thread
Numeric relation extraction
Knowledge Graph over a thread
Source Credibility
Aspect-based Sentiment
Information Distortion

RQ 1
What effects do differences in the source credibility of debunking messages have on WhatsApp conversations

RQ 2
How do sentiments change in response to debunking messages?

RQ 3
How does information distortion change in response to debunking messages

Figure 1. Process flow diagram of this study: It outlines how debunking threads are extracted from the WhatsApp messages and how the debunking messages are used for analysis

Topic classification

Firstly, *top2vec* (Angelov, 2020), an unsupervised topic modelling technique, was used to identify the most representative topics using keywords and tagging each WhatsApp message to a topic. Recently, *top2vec* was used for topic modelling of COVID-19 news (Ghasiya and Okamura, 2021). This can yield more relevant results over traditional methods like Latent Dirichlet Allocation (LDA), as *top2vec* can automatically generate the number of topics, whereas LDA requires the number of topics to be pre-specified.

The entire corpus used for TC on the data set comprises 78,921 sentences. A sample of the *top2vec* topic model results was manually annotated to construct a gold standard training dataset (4,477 sentences and 62 topic labels) for TC.

A pre-trained BERT model (*bert-base-case*) was fine-tuned on the gold data set for TC. The data set was split into 80% training and 20% test data. The test accuracy of the model is 0.80. The dropout probability was set at 0.2, and the number of epochs to 48, and the initial learning rate was 1e-5, batch size 5. The sentences of all messages were classified using the BERT-based topic classifier.

Coders A and B sampled 5% of the predicted topics (200 messages) to check the output quality of the topic classifier. For intercoder reliability analysis between the coders and the classifier's topic label (Artstein and Poesio, 2008), Fleiss's kappa (κ) was acceptable (κ = 0.86).

Extracting a debunking message thread

Extracting a debunking message thread involves manual identification of a debunking thread which comprises at least one instance of an IDO/a reference to an IDO and one instance of a debunking message within a conversation thread (see Figure 1).

The maximum thread time between messages is capped at two days, and a message thread is classified as a new message thread when the topic has distinctly shifted. We chose this time on the basis of 1) there were calls to reconvene discussions to let people rest at night, 2) a conversation thread can be derailed and revisited later and 3) literature suggesting that group conversations can last for 50 h (Caetano *et al.*, 2018).

Finally, using input from the topic classifier for reference, each debunking thread was manually labelled with a topic label by Coders A and B. Agreement is 94%, and Cohen's kappa (Artstein and Poesio, 2008) was acceptable (κ = 0.93). All messages retained for analysis were selected through a majority vote, where coder C acted as a tiebreaker.

Handling of multimedia messages

All the available multimedia (e.g. fake videos of Singaporeans raiding a supermarket) in the messages were manually labelled. However, the multimedia data set was partially complete due to various reasons such as deletion, or the multimedia is not available for download. If the multimedia is absent or irrelevant to a debunking thread, it is excluded from the analysis.

Categorising different types of debunking messages and URL

The debunking messages were categorised into four types: doubt veracity, organic, link and official communications (see Table 1) through a ground-up examination of debunking messages. All URLs were categorised into various tiers based on their domain names.

Aspect-term sentiment analysis

To classify sentiment towards aspects in the message, an aspect-term SA (ATSA) approach was used. ATSA is a variation of ABSA, where the explicit aspect-term rather than aspect categories are used for sentiment classification. For instance, in the phrase, “I love the phone screen and the OS”, the explicit aspect-term “phone screen” and “OS” are extracted, while the sentiment for both is positive. Following the approach by Sun *et al.* (2019), ATSA is treated as a sentence-pair classification task that comprises a sentence, an aspect-term and the sentiment label. For aspect-term extraction, various scholars used different methods like topic models or neural networks. However, Tulkens and van Cranenburgh (2020) found a more straightforward approach using word embeddings and POS taggings. Hence, a similar approach was adopted by performing aspect-term extraction on the sentences using named entity and noun extraction using *stanza* (Qi *et al*., 2020).

In order to construct a gold standard training data set, a pre-trained RoBERTa model trained on ~58 M tweets (Barbieri *et al*., 2020) was used to tag sentences from the WhatsApp data set with predicted sentiment labels. Sentences with the relevant aspect-terms and correct sentiment labels were then selected to create the training dataset (5,989 sentence-aspect-sentiment triples that were not in the debunking threads). The training dataset was used to train a BERT-based ATSA classifier with a good test *f1* of 0.93.

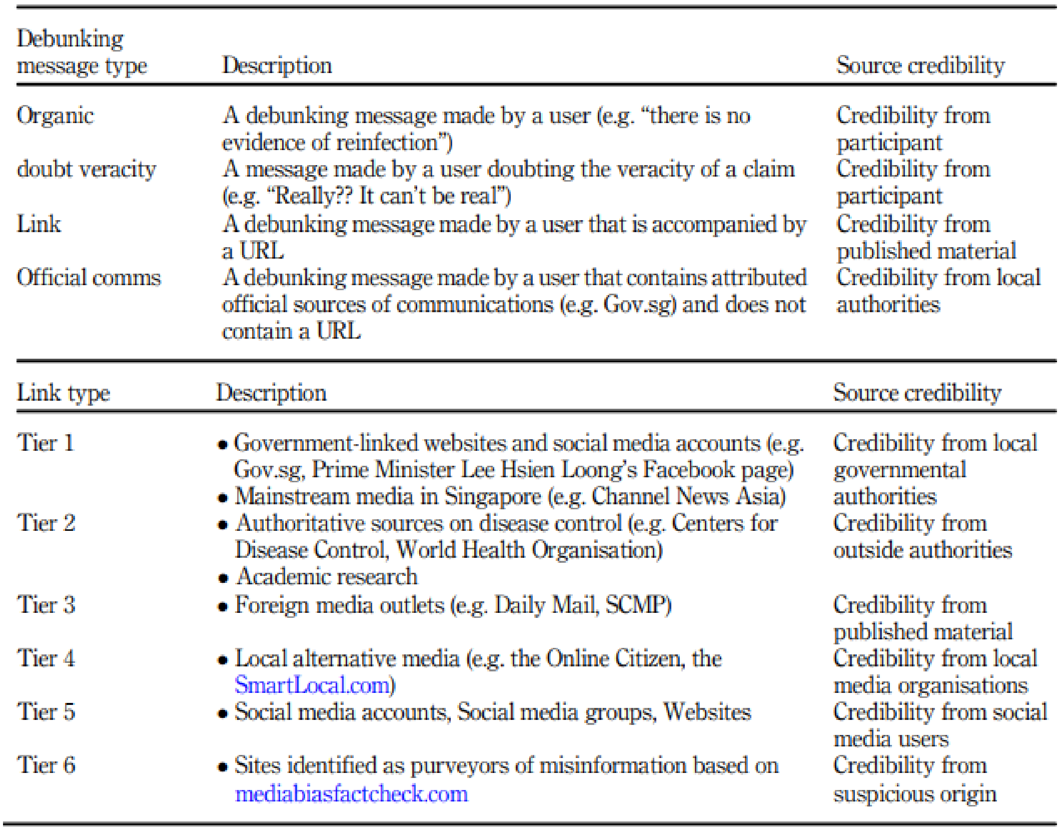
Next, aspect-term extraction for all sentences within the debunking threads was done to identify relevant aspect-sentence pairs. Additionally, pronoun resolution was manually done to capture aspect-terms more precisely. Afterwards, the BERT-based ATSA classifier was used to predict labels in the data set of debunking threads (4,070 sentence-aspect pairs). Finally, the aspect-terms were regrouped into 27 aspect categories using a ground-up approach.

Coders A and B sampled 5% of the predicted data (200 messages) to check the ATSA classifier's output quality. For intercoder reliability analysis between the coders and classifier's sentiment label (Artstein and Poesio, 2008), Fleiss's kappa (κ) was acceptable (κ = 0.83).

Analysing of pre-post sentiment towards aspects

To analyse pre-post sentiment towards an aspect category, the following was done: Given a debunking message thread, *DThread*, containing *n* number of messages *m*, $DThread = \{m_1 + m_2 \ldots + m_k \ldots + m_n\}$, the position of each message *mPos* is rescaled by dividing the message position *k* over the number of messages $mPos_k = k/n$.

Next, message *m* was split into *j* sentences $\{\text{sentence}_1 + \text{sentence}_2 \ldots + \text{sentence}_j\}$, which contains the following features: 1) $mPos_k$, 2) *aspect* at the sentence level and 3) sentiment polarity. Finally, the mean sentiment of an aspect pre and post debunking message was used for statistical analysis.



Debunking message type	Description	Source credibility
Organic	A debunking message made by a user (e.g. "there is no evidence of reinfection")	Credibility from participant
doubt veracity	A message made by a user doubting the veracity of a claim (e.g. "Really?? It can't be real")	Credibility from participant
Link	A debunking message made by a user that is accompanied by a URL	Credibility from published material
Official comms	A debunking message made by a user that contains attributed official sources of communications (e.g. Gov.sg) and does not contain a URL	Credibility from local authorities

Link type	Description	Source credibility
Tier 1	• Government-linked websites and social media accounts (e.g. Gov.sg, Prime Minister Lee Hsien Loong's Facebook page) • Mainstream media in Singapore (e.g. Channel News Asia)	Credibility from local governmental authorities
Tier 2	• Authoritative sources on disease control (e.g. Centers for Disease Control, World Health Organisation) • Academic research	Credibility from outside authorities
Tier 3	• Foreign media outlets (e.g. Daily Mail, SCMP)	Credibility from published material
Tier 4	• Local alternative media (e.g. the Online Citizen, the SmartLocal.com)	Credibility from local media organisations
Tier 5	• Social media accounts, Social media groups, Websites	Credibility from social media users
Tier 6	• Sites identified as purveyors of misinformation based on mediabiasfactcheck.com	Credibility from suspicious origin

Table 1. Different types of debunking messages and link types

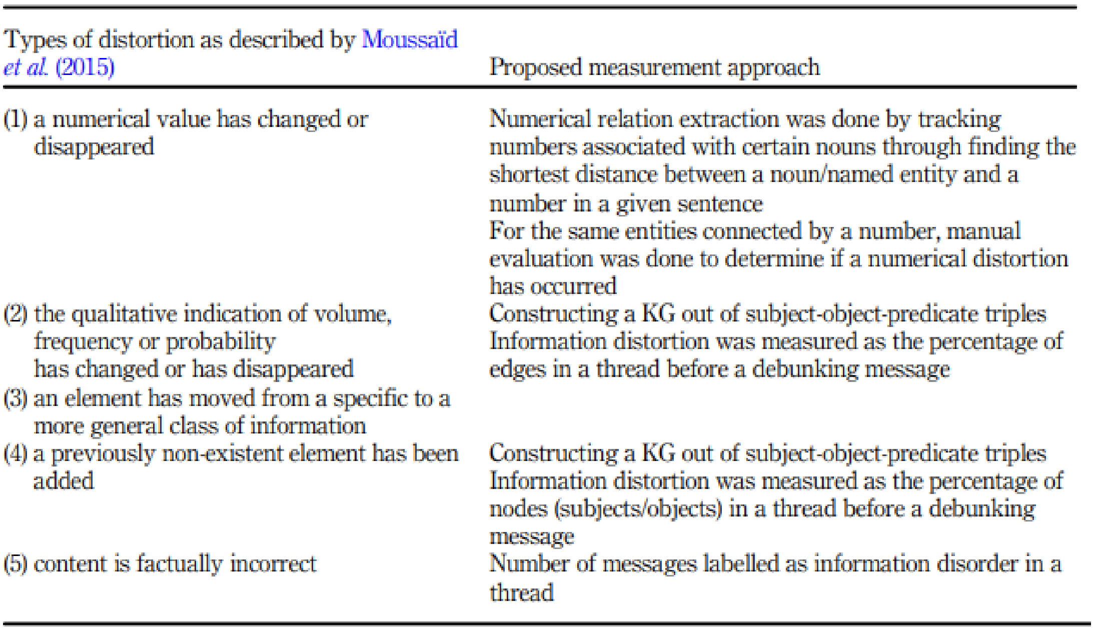
Analysing the impact of the debunking message on the discussion

To analyse the impact of the debunking message on discussion thread length, *mPos* of all types of debunking messages in the *DThread* were used to construct an empirical cumulative distribution function (ECDF).

Measuring information distortion

Moussaïd *et al.* (2015) manually coded IDT using the following types of distortions: (1) a numerical value has changed or disappeared; (2) a qualitative indication of volume, frequency or probability has changed or disappeared; (3) an element has moved from a specific to a more general class of information; (4) a previously non-existent element has been added; and (5) content is wrong.

In Moussaïd *et al.*'s (2015) coding scheme, IDT is measured by changes in one message as it passes through different recipients. However, this approach is not suited for WhatsApp group chats as they are conversations where people contribute new information, give irrelevant comments or even share IDO. Hence, the IDT typology codebook was adapted (see Table 2) using IDO labels, KG and numerical relation extraction to identify IDT consistently.



Types of distortion as described by Moussaïd *et al.* (2015)	Proposed measurement approach
(1) a numerical value has changed or disappeared	Numerical relation extraction was done by tracking numbers associated with certain nouns through finding the shortest distance between a noun/named entity and a number in a given sentence For the same entities connected by a number, manual evaluation was done to determine if a numerical distortion has occurred
(2) the qualitative indication of volume, frequency or probability has changed or has disappeared	Constructing a KG out of subject-object-predicate triples Information distortion was measured as the percentage of edges in a thread before a debunking message
(3) an element has moved from a specific to a more general class of information	
(4) a previously non-existent element has been added	Constructing a KG out of subject-object-predicate triples Information distortion was measured as the percentage of nodes (subjects/objects) in a thread before a debunking message
(5) content is factually incorrect	Number of messages labelled as information disorder in a thread

Table 2. Summary of adaptions to the information distortion typology codebook

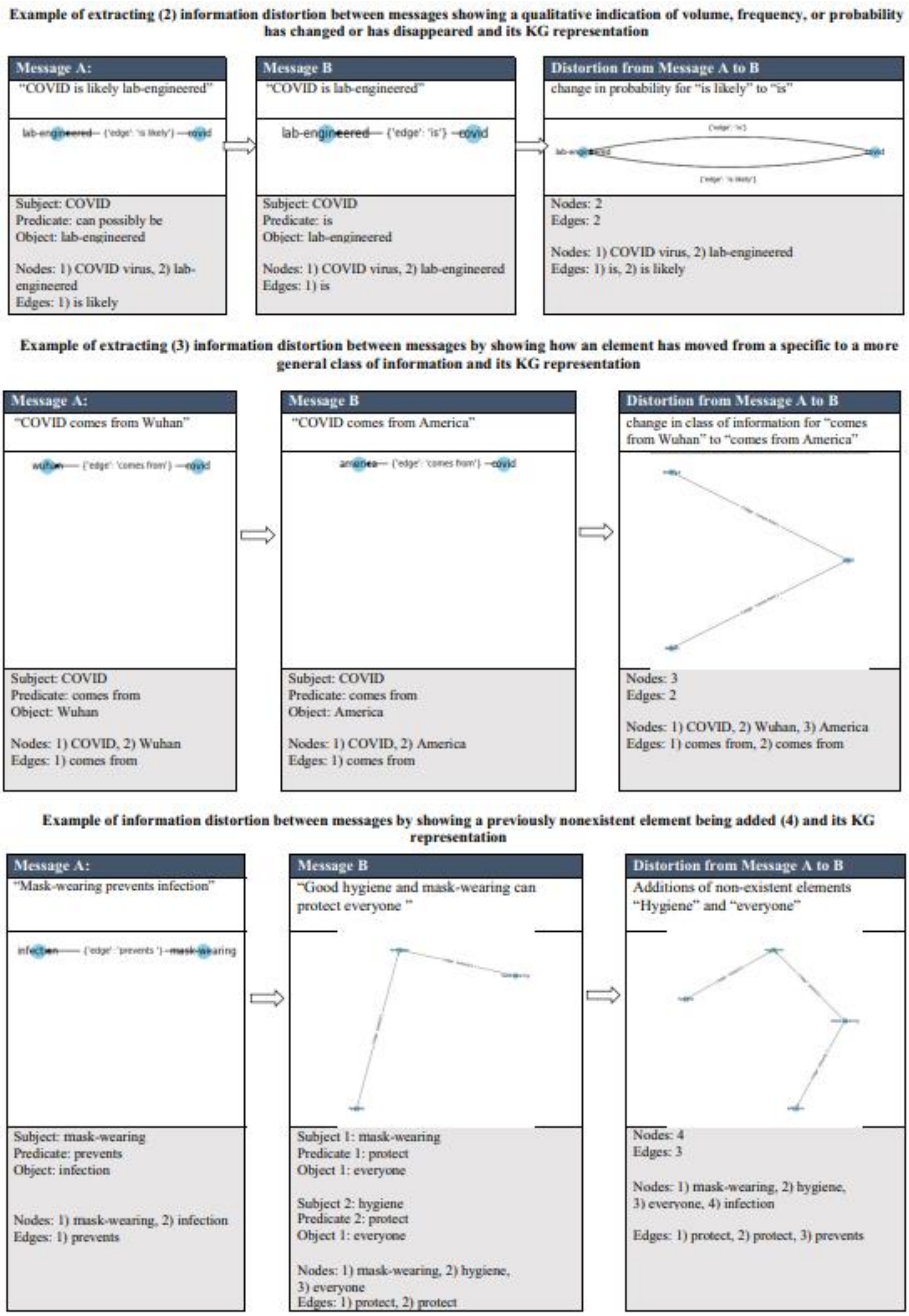
Numeric relation extraction

For (1), the numerical relation extraction approach by Madaan *et al.* (2016) was employed. A dependency parse is created using NLTK's POS tagger for every sentence. Next, the shortest path dependency parse between a named entity/noun and a number is identified as the relation between the number and entity.

Coders A and B reviewed the extracted number-entity pairs to determine if a numerical distortion has occurred (e.g. one message claimed that there are 10,000 people from the Hubei province currently in Singapore, while another claimed that the number is 2,000). Cohen's kappa (κ) was acceptable ($\kappa = 0.88$). Only agreements were retained for analysis (three messages were excluded from analysis as a result).

Knowledge graph construction

To extract (2), (3), (4), *stanza* (Qi *et al.*, 2020) was used to extract open-domain relation triples consisting of a subject-predicate-object. A knowledge graph (KG) was constructed from these triplets using *networkx* (Hagberg *et al.*, 2008), which grows with every message containing valid triplets adding to the KG (see Figure 2).



Example of extracting (2) information distortion between messages showing a qualitative indication of volume, frequency, or probability has changed or has disappeared and its KG representation

Message A:	Message B	Distortion from Message A to B
"COVID is likely lab-engineered"	"COVID is lab-engineered"	change in probability for "is likely" to "is"
Subject: COVID Predicate: can possibly be Object: lab-engineered Nodes: 1) COVID virus, 2) lab-engineered Edges: 1) is likely	Subject: COVID Predicate: is Object: lab-engineered Nodes: 1) COVID virus, 2) lab-engineered Edges: 1) is	Nodes: 2 Edges: 2 Nodes: 1) COVID virus, 2) lab-engineered Edges: 1) is, 2) is likely

Example of extracting (3) information distortion between messages by showing how an element has moved from a specific to a more general class of information and its KG representation

Message A:	Message B	Distortion from Message A to B
"COVID comes from Wuhan"	"COVID comes from America"	change in class of information for "comes from Wuhan" to "comes from America"
Subject: COVID Predicate: comes from Object: Wuhan Nodes: 1) COVID, 2) Wuhan Edges: 1) comes from	Subject: COVID Predicate: comes from Object: America Nodes: 1) COVID, 2) America Edges: 1) comes from	Nodes: 3 Edges: 2 Nodes: 1) COVID, 2) Wuhan, 3) America Edges: 1) comes from, 2) comes from

Example of information distortion between messages by showing a previously nonexistent element being added (4) and its KG representation

Message A:	Message B	Distortion from Message A to B
"Mask-wearing prevents infection"	"Good hygiene and mask-wearing can protect everyone "	Additions of non-existent elements "Hygiene" and "everyone"
Subject: mask-wearing Predicate: prevents Object: infection Nodes: 1) mask-wearing, 2) infection Edges: 1) prevents	Subject 1: mask-wearing Predicate 1: protect Object 1: everyone Subject 2: hygiene Predicate 2: protect Object 1: everyone Nodes: 1) mask-wearing, 2) hygiene, 3) everyone Edges: 1) protect, 2) protect	Nodes: 4 Edges: 3 Nodes: 1) mask-wearing, 2) hygiene, 3) everyone, 4) infection Edges: 1) protect, 2) protect, 3) prevents

Figure 2. Extracting (2), (3), (4) from Moussaïd *et al.* (2015)'s typology

For (2) and (3), IDT was measured by tracking predicate changes. The KG approach can represent an increase in edges from message to message as an increase in new predicates in a KG. For (4), the addition of non-existent elements can be represented as an increase in the number of new nodes in a KG.

Information disorder labels
To extract (5), the number and positions of messages labelled as IDO were tracked, which was done during the extraction of debunking messages.

Results

Types of debunking messages
There were 102 debunking threads (2,649 messages, 104 unique user ids) identified. 74.7% of debunking messages were organic and doubt veracity. In comparison, 25.3% of debunking messages were links and official messages.

Most links were authoritative ones, as 88.4% of links used in debunking were tier 1 and tier 2 links. None of the local alternative media or dubious sites was used in debunking threads. There was only one Tier 5 link, making it difficult to do any meaningful analysis. Hence, Tier 1, 2 and 3 links were used for analysis.

Effect of debunking messages on length of the conversation
For RQ1, debunking messages were compared based on the ECDF (see Figure 3). It indicated that official communications and links (50% of these messages are found after 0.64 position in the thread) were more likely to be found later in a thread than organic debunking or doubt-veracity messages (50% of these messages are found after 0.5 and 0.35 position in the thread, respectively).

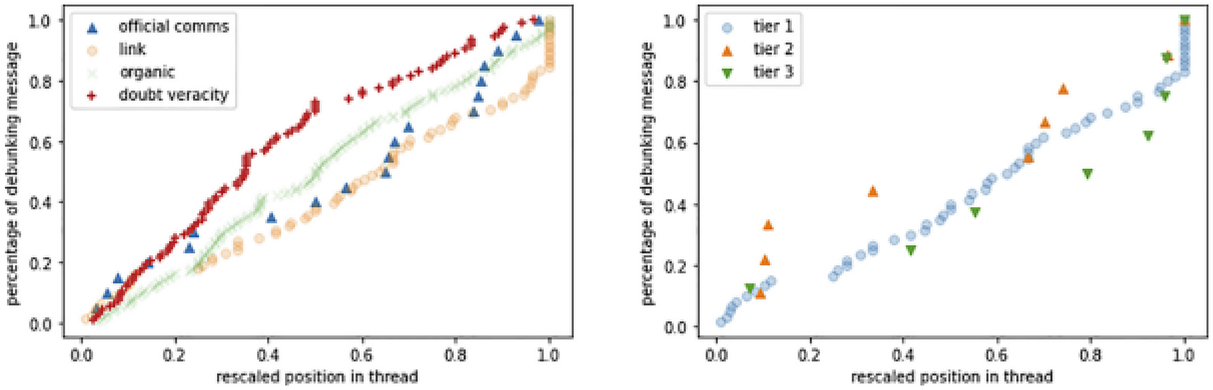


Figure 3. ECDF analysis of the positions of the debunking messages in a thread (left) and the analysis of the different tier levels of the links (right). The *X*-axis shows all messages in a thread that are rescaled to be between 0 and 1. The *Y*-axis shows the cumulative percentage of the debunking messages for each type

ECDF analysis for types of debunking links indicated that tier 2 and tier 3 sources were more likely to be found later in a thread than other sources (i.e. 50% of these links are found after 0.66 and 0.85 positions in the thread, respectively). Tier 1 sources tend to be found earlier in the thread (50% of these links are found after 0.62 position in the thread), indicating that sharing links from these sources is likely to result in further discussion. It is possible that information from government sources or local media were more relevant to the local public, hence generating more discussion than external information.

Comparing 1) the percentage of messages in a thread before a debunking message and 2) time elapsed for messages in a thread before a debunking message indicated that low source credibility debunking messages were likely to be found earlier in a thread. There was a statistically significant difference between debunking messages for the percentage of messages before a debunking message as determined by Kruskal–Wallis one-way analysis of variance (ANOVA) ($H(3, 193) = 19.5$, $p < 0.01$). A pairwise post-hoc Dunn test showed that doubt veracity-type debunking message was significantly earlier in the thread than organic and link types ($p < 0.05$).

Similarly, there was a statistically significant difference between debunking messages in terms of the time elapsed for messages in a thread before a debunking message as determined by Kruskal-Wallis one-way

ANOVA (H(3,193) = 16.98, $p < 0.01$). A pairwise post-hoc Dunn test showed that timing elapsed for doubt veracity-type debunking message was significantly earlier in the thread than all types ($p < 0.05$).

For both percentage of messages and time elapsed, there was no statistically significant difference between tier 1, 2 and 3 debunking links as determined by Kruskal-Wallis one-way ANOVA.

Effects of debunking messages on sentiment

For RQ2, sentiments towards various aspects discussed in the debunking threads were identified. 4,461 sentence-aspect pairs and 27 aspects categories were analysed. In general, sentiments in debunking threads were neutral (58.1%), followed by negative sentiment (30.1%) and positive (11.8%).

Aspects with the top three largest proportions of negative sentiments were Singapore–Malaysia's relations, coronavirus-related information and public transport (see Figure 4). Conversely, aspects with the top 3 largest proportion of positive sentiments were dubious advice for protecting against COVID-19, hygiene advice and racism and COVID-19 immunity.

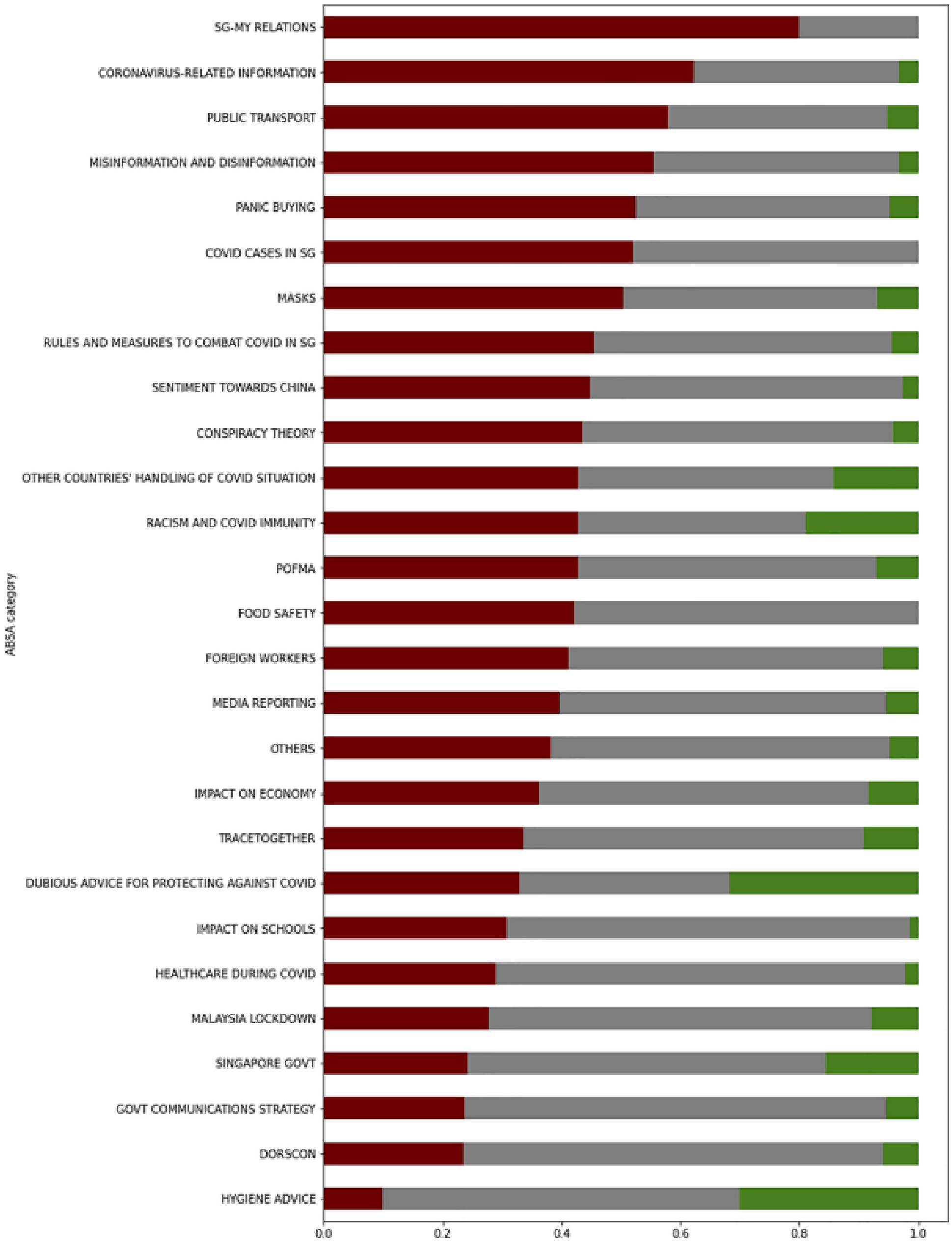


Figure 4 Percentage of sentiments across various aspects

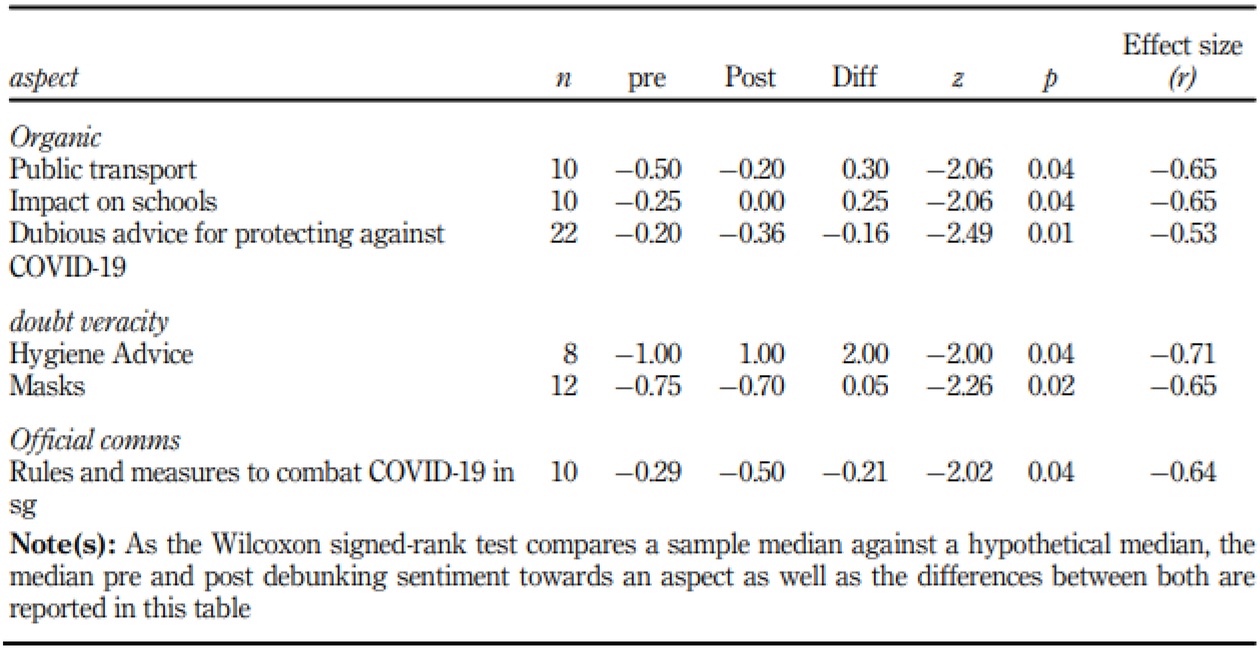
Support for hygiene measures could be attributed to positive attitudes towards these measures in protecting oneself from COVID-19. However, not all positive sentiments towards an aspect are desirable. For instance, when sharing dubious advice for protecting oneself against COVID-19, participants may be enthusiastic about untested remedies against COVID-19. For cases involving racism, lack of COVID-19 cases amongst certain ethnic groups (due to the early stages of the pandemic) can trigger unscientific beliefs that these groups are, by nature, immune to the virus.

Negative attitudes can be attributed to worries about the outbreaks, such as crowding on public transportation or worries about an uncontrollable outbreak. Concerns about Singapore–Malaysia's relations were due to the spread of COVID-19 in Malaysia, leading to the MCO. It triggered negative attitudes towards Malaysia's MCO, as it affected neighbours like Singapore.

Changes in sentiments towards aspects pre- and post-debunking

Based on the Wilcoxon signed-rank test (Woolson, 2008), there was no significant difference in general sentiments pre and post debunking messages for debunking threads, whether for organic, doubt-veracity, link or official communications. This was surprising as other rumour research indicated that negative sentiments generally improved after refuting these rumours (Zeng and Zhu, 2019).

However, there were statistically significant differences for the following aspects pre and post debunking message (see Table 3) for six aspects of this COVID-19 situation: *impact on schools, dubious advice for protecting against COVID-19, public transport, masks, hygiene advice and rules and measures to combat COVID-19*.



aspect	*n*	pre	Post	Diff	*z*	*p*	Effect size *(r)*
Organic							
Public transport	10	−0.50	−0.20	0.30	−2.06	0.04	−0.65
Impact on schools	10	−0.25	0.00	0.25	−2.06	0.04	−0.65
Dubious advice for protecting against COVID-19	22	−0.20	−0.36	−0.16	−2.49	0.01	−0.53
doubt veracity							
Hygiene Advice	8	−1.00	1.00	2.00	−2.00	0.04	−0.71
Masks	12	−0.75	−0.70	0.05	−2.26	0.02	−0.65
Official comms							
Rules and measures to combat COVID-19 in sg	10	−0.29	−0.50	−0.21	−2.02	0.04	−0.64

Note(s): As the Wilcoxon signed-rank test compares a sample median against a hypothetical median, the median pre and post debunking sentiment towards an aspect as well as the differences between both are reported in this table

Table 3. Significant differences in sentiment towards various aspects pre and post debunking messages

The change in sentiment for each aspect in a thread was measured for all debunking types: organic (22 aspects), doubt-veracity (20 aspects), link (20 aspects) and official communications (14 aspects). Except for links, other debunking messages have aspects with significant differences in sentiment pre and post debunking. One plausible explanation is that out of 78 links, 42 of them are posted without accompanying text. The lack of text could be a missing cue that influenced people's sentiment post-debunking.

Results and findings on information distortion in a debunking thread

For RQ3, IDT throughout a thread was tracked to see how IDT accumulates and how different debunking messages affect it. Different types of IDT were compared based on the ECDF (see Figure 5).

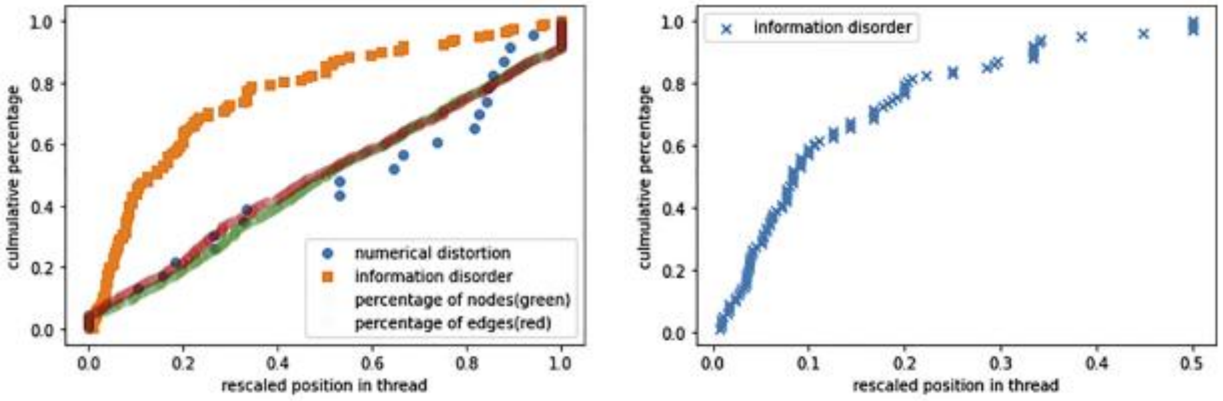


Figure 5. ECDF analysis of messages with IDT and their positions in a given thread (left) and the ECDF of the positions of the first IDT in all debunking threads (right)

Messages that carry information disorder

Messages that carry IDO were more likely to be found early in the debunking thread. For example, more than half of the messages labelled as IDO were found earlier in the thread compared to other types of IDT. Furthermore, for the first instance of an IDO in every debunking thread, 80% was found early in a given thread (position 0.0–0.2). None were found halfway in a thread (position > 0.6).

Changes or disappearance in numerical value

Within 11 threads, there were 23 messages where the numerical value had changed or disappeared. However, there appeared to be no specific pattern in the position of or the number of numerical distortions in a debunking thread.

Addition of a previously non-existent element and predicates

Comparisons of the addition of non-existent elements after a debunking message were done across all debunking types. A Kruskal–Wallis one-way ANOVA test (McKight and Najab, 2010) showed that there was a statistically significant difference between debunking messages ($H(3,378) = 23.1$, $p < 0.01$). A pairwise post-hoc Dunn test showed significantly more new nodes after doubt veracity-type debunking messages than all other types of debunking messages ($p < 0.05$).

Comparisons of the addition of more predicates after a debunking message were done across all debunking types. There was a statistically significant difference between debunking messages as determined by Kruskal–Wallis one-way ANOVA ($H(3,378) = 20.9$, $p < 0.01$). A pairwise post-hoc Dunn test showed significantly more predicates added after doubt veracity-type debunking messages than all other types of debunking messages.

Comparisons of 1) the addition of more predicates after a debunking message as well as 2) the addition of non-existent elements after a debunking message were made across different types of links within the link category. There was no statistically significant difference between tier 1, 2 and 3 debunking links as determined by Kruskal–Wallis one-way ANOVA.

Discussion

Effect of debunking messages on conversation length

For RQ1, the position of debunking messages, percentages of messages and time elapsed before debunking elapsed all point to a consistent observation. Sources with higher credibility (i.e. link and official communications) were more likely to be found later in a message thread than sources with lower credibility (i.e. organic and doubt veracity). As debunking messages with higher source credibility were likely to be associated with being found in a later position in any given thread, it suggests that source credibility have some impact on closing a discussion thread and getting participants to move to a different topic.

Additionally, the lag for higher source credibility could be that quality fact-checks take time. During health crises, giving inconsistent or unclear information can adversely impact the public's decision to adopt behaviours that can help prevent an outbreak (Lundgren and McMakin, 2013). Hence, agencies need to be timely and careful in debunking false information, explaining the lag.

Debunking messages with high source credibility are associated with earlier conversation-stopping than debunking messages drawn from individuals' judgement. This extends prior work by Ermakova *et al.* (2020), who found that debunkers who were seen as authoritative sources of information on COVID-19 issues cut short information cascades relating to the virus. The current study found that this earlier stopping effect is not limited to debunkers but also to messages linked/referenced to authoritative sources.

Sentiment changes pre- and post-debunking

For RQ2, findings indicated that sentiments towards various aspects of a crisis could change bidirectionally in response to debunking messages – a relatively unexplored area in the literature.

Sentiment towards *impact on schools*, *public transport*, *masks* and *hygiene advice* significantly improved after debunking. Based on a qualitative dive into the data, improvement in sentiments for impact on schools could be attributed to the debunking of rumours of school closures, which is distressing for parents. Similarly, improvement in sentiment towards public transportation matters could be attributed to the correcting rumours that all bus services were suspended. This is consistent with the current understanding in the literature, where negative sentiments generally improved after these rumours were refuted (Zeng and Zhu, 2019).

For *hygiene advice*, improvement in sentiment could be attributed to the promotion of handwashing as vital advice to protect oneself from COVID-19. For *masks*, the improvement could be linked to participants sharing advice to address wrong methods of mask-wearing.

In contrast, for *rules and measures to combat COVID-19*, deterioration in sentiment can be attributed to fears following the announcement of new COVID-19 measures. For example, false information about a pending lockdown was criticised as being recycled fake news about rules and measures to combat COVID-19.

For *dubious advice for protecting against COVID-19*, deterioration of sentiment could be attributed to post-debunking criticisms directed against dubious cures like essential oils. It shows that negative sentiment towards certain aspects could be welcome when directed towards dubious cures or fake claims.

Although this study found only a few aspects that significantly changed after a debunking message, ABSA could be valuable for public agencies in evaluating their efforts in countering COVID-19 IDO. In practice, the findings can act as a sense-making signal for public agencies monitoring public sentiments towards various aspects of COVID-19.

Information distortion

The longer a conversation thread about a piece of IDO is the more distortion is likely to happen, reinforcing the IDO. Debunking could cut short such conversations, as debunkers furnish countering evidence that becomes accepted, and the conversation would move to other topics (see Ermakova *et al.*, 2020). Hence, for RQ3, IDT was tracked throughout a thread to measure the effect of debunking messages.

A total of four approaches were taken to measure IDT: Numerical relation extraction to identify numerical distortions, growth in nodes and edges in a KG of the debunking thread, and the number of messages labelled as IDO. This is a novel approach to measuring IDT by adapting the IDT typology codebook proposed by Moussaïd *et al.* (2015) to perform analysis in a scalable fashion on observational data. It is also a radical shift from previous methods (Moussaïd *et al.*, 2015; Carlson, 2017; Ribeiro *et al.*, 2019), which relied on experimental research designs where participants iteratively summarised a text to track the IDT.

Analysis of KGs supported the idea that source credibility of debunking is relevant in reducing IDT. Doubt veracity-type debunking messages were more likely associated with the addition of non-existent elements or more predicates after the message than other types of messages. It is possible that the simple quality of doubt-veracity type debunking message was unable to address the misinformation discussed, consequently sparking more discussion. Related research on debunking supports this inference as simple corrections are unlikely to affect beliefs in misinformation (Lewandowsky *et al.*, 2021).

IDO messages were found early in the debunking thread. It was expected that in a debunking thread, the IDO should appear early in a thread for debunking to occur.

Although there is no specific pattern found for numerical distortions, a qualitative examination indicated that numerical distortions were either: 1) inconsistent numbers for things like the number of days of medical leave or 2) numbers about certain things or events that were wrong such as the number of deaths or an inaccurate timing.

In practice, the results highlight the importance of crafting debunking messages attributable to authoritative sources and designed for WhatsApp contexts. Source credibility can reduce IDT on WhatsApp chats. This is aligned with current advice on the role of credible sources such as news media or health agencies in refuting false information (Lewandowsky *et al.*, 2021). The results also suggest that the sooner the authorities debunk the misinformation, the less likely and severe the IDT.

Conclusion

First, the descriptive analysis, ECDF analysis, and comparison of time elapsed corroborate the understanding that debunking messages with higher source credibility are associated with the earlier stopping of conversations than debunking messages drawn from the individual's judgement. Secondly, debunking messages can have bidirectional effects on sentiments as they can significantly improve sentiments towards some aspects of the COVID-19 situation, such as its impact on schools, public transport, masks and hygiene advice, while significantly deteriorating sentiments towards rules and measures to combat COVID-19 and dubious advice for protecting against COVID-19 in Singapore. Finally, source credibility is vital in reducing IDT on WhatsApp conversations. Analysis of KGs showed that doubt veracity-type debunking messages were more likely to be associated with the addition of non-existent elements or more predicates after the message than other types of messages.

Due to the exploratory nature of this research, there were some limitations. Firstly, the findings might not be transferable as it studies a single WhatsApp group chat's discussion on the early part of the COVID-19 pandemic in Singapore. The behaviour and patterns observed could be influenced by the discussion topic. For example, the results for a COVID-19 pandemic thread may differ from a political chat group thread. Nonetheless, this study has provided results based on a substantial WhatsApp data set, which provides a springboard for future studies. Next, the current KG approach tracks the morphing of information throughout a conversation thread to indicate IDT. Future work can identify differences within types of additions of nodes and predicates to refine the analysis. Additionally, the current measurement of predicates combined the analysis for 1) qualitative indication of volume, frequency or probability has changed or has disappeared and 2) an element has moved from a specific to a more general class of information. Future work could disentangle that.

These limitations notwithstanding, this exploratory study aimed to understand online responses to debunking messages about COVID-19 in Singapore and present the following contributions: Firstly, it found support for the importance of source credibility in debunking and the practical role of an aspect-based approach in analysing sentiment changes in evaluating responses to debunking messages. These findings have practical value for public agencies during a health crisis. Additionally, these results provide a significant first step towards examining the effects of debunking on sentiment at the aspect level. Next, WhatsApp conversations were studied, which departs from the approaches in existing literature, which were either experimental or done entirely on more open social media platforms like Twitter. Finally, novel approaches were used to perform content analysis on a COVID-19-related public chat on WhatsApp, like fine-tuning deep learning techniques BERT for aspect-term SA and TC. Novel approaches for the study of IDT by using KGs to measure IDT were used.

References

Abdul Rahman, A. (2020), “Mitigating the social pandemic of xenophobia during COVID-19”, in Khader, M., Dillon, D., Chen, X.K., Neo, L.S. and Chin, J. (Eds), How to Prepare for the Next Pandemic: Behavioural Sciences Insights for Practitioners and Policymakers, World Scientific Press.

Aggarwal, A., Chauhan, A., Kumar, D., Mittal, M. and Verma, S. (2020), “*Classification of fake news by fine-tuning deep bidirectional Transformers based language model*”, EAI Endorsed Transactions on Scalable Information Systems, Vol. 7 No. 27, doi: 10.4108/eai.13-7-2018.163973.

Alamoodi, A.H., Zaidan, B.B., Zaidan, A.A., Albahri, O.S., Mohammed, K.I., Malik, R.Q., Almahdi, E.M., Chyad, M.A., Tareq, Z., Albahri, A.S., Hameed, H. and Alaa, M. (2021), “*Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: a systematic review*”, Expert Systems with Applications, Vol. 167, p. 114155, doi: 10.1016/j.eswa.2020.114155.

Angelov, D. (2020), “*Top2Vec: distributed representations of topics*”, arXiv, available at: http://arxiv.org/abs/2008.09470 (accessed 25 January 2021).

Apuke, O.D. and Omar, B. (2020), “*User motivation in fake news sharing during the COVID-19 pandemic: an application of the uses and gratification theory*”, Online Information Review, Vol. 45 No. 1, pp. 220-239, doi: 10.1108/OIR-03-2020-0116.

Artstein, R. and Poesio, M. (2008), “*Inter-coder agreement for computational linguistics*”, Computational Linguistics, Vol. 34 No. 4, pp. 555-596, doi: 10.1162/coli.07-034-R2.

Ayoub, J., Yang, X.J. and Zhou, F. (2021), “*Combat COVID-19 infodemic using explainable natural language processing models*”, Information Processing and Management, Vol. 58 No. 4, p. 102569, doi: 10.1016/j.ipm.2021.102569.

Barbieri, F., Camacho-Collados, J., Neves, L. and Espinosa-Anke, L. (2020), TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, *arXiv:2010.12421 [cs]*, available at: http://arxiv.org/abs/2010.12421 (accessed 9 March 2021).

Barbosa, S. and Milan, S. (2019), “*Do not harm in private chat apps: ethical issues for research on and with WhatsApp*”, Westminster Papers in Communication and Culture, Vol. 14.

Baruah, A., Das, K., Barbhuiya, F. and Dey, K. (2020), Automatic Detection of Fake News Spreaders Using BERT, CLEF.

Basu, M. (2020), “*Exclusive: how Singapore sends daily WhatsApp updates on coronavirus*”, GovInsider, available at: https://govinsider.asia/innovation/singapore-coronavirus-whatsapp-covid19-open-government-products-govtech/ (accessed 12 April 2021).

Bordia, P., DiFonzo, N., Haines, R. and Chaseling, E. (2005), “*Rumors denials as persuasive messages: effects of personal relevance, source, and message characteristics*”, Journal of Applied Social Psychology, Vol. 35 No. 6, pp. 1301-1331, doi: 10.1111/j.1559-1816.2005.tb02172.x.

Bowles, J., Larreguy, H. and Liu, S. (2020), “*Countering misinformation via WhatsApp: evidence from the COVID-19 pandemic in Zimbabwe*”, CID Working Paper Series, available at: https://dash-harvard-edu.libproxy.smu.edu.sg/handle/1/37366416 (accessed 01 December 2021).

Bursztyn, V.S. and Birnbaum, L. (2019), “*Thousands of small, constant rallies: a large-scale analysis of partisan WhatsApp groups*”, 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 484-488, doi: 10.1145/3341161.3342905.

Caetano, J.A., de Oliveira, J.F., Lima, H.S., Marques-Neto, H.T., Magno, G., Meira, W. Jr, Almeida, V.A. (2018), Analyzing and Characterizing Political Discussions in WhatsApp Public Groups, arXiv preprint arXiv:1804.00397.

Carlson, T.N. (2017), “*Modeling political information transmission as a game of telephone*”, The Journal of Politics, Vol. 80 No. 1, pp. 348-352, doi: 10.1086/694767.

Chong, M. and Choy, M. (2018), “*The social amplification of haze-related risks on the internet*”, Health Communication, Vol. 33 No. 1, pp. 14-21, doi: 10.1080/10410236.2016.1242031.

Chan, M.S., Jones, C.R., Hall Jamieson, K. and Albarracín, D. (2017), “*Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation*”, Psychological Science, Vol. 28 No. 11, pp. 1531-1546.

Chong, C. (2020), “*Coronavirus: govt debunks fake news on Singaporeans contracting the virus and Singapore running out of masks*”, The Straits Times, 31 January, available at: https://www.straitstimes.com/singapore/wuhan-virus-govt-debunks-fake-news-on-singaporeans-contracting-the-virus-singapore-running (accessed 11 July 2021).

Coombs, W.T. (2015), “*The value of communication during a crisis: insights from strategic communication research*”, Business Horizons, Vol. 58 No. 2, pp. 141-148.

CoronaVirusFacts Alliance (no date), “*Poynter*”, available at: https://www.poynter.org/coronavirusfactsalliance/ (accessed 15 June 2021).

COVID-19 global risk communication and community engagement strategy, December 2020-May 2021: interim guidance, 23 December 2020 (2020), World Health Organization.

de Souza, J.V., Gomes, J., de Souza Filho, F.M., de Julio, A.M. and de Souza, J.F. (2020), “*A systematic mapping on automatic classification of fake news in social media*”, Social Network Analysis and Mining, Vol. 10 No. 1, pp. 1-21, doi: 10.1007/s13278-020-00659-2.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018), *Bert: Pre-training of Deep Bidirectional Transformers for* Language Understanding, arXiv.

Do, H.H., Prasad, P., Maag, A. and Alsadoon, A. (2019), “*Deep learning for aspect-based sentiment analysis: a comparative review*”, Expert Systems with Applications, Vol. 118, pp. 272-299, doi: 10.1016/j.eswa.2018.10.003.

Ermakova, L., Nurbakova, D. and Ovchinnikova, I. (2020), “*Covid or not covid? Topic shift in information cascades on twitter*”, Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM), COLING-RDSM 2020, Barcelona, Association for Computational Linguistics, pp. 32-37, available at: https://www.aclweb.org/anthology/2020.rdsm-1.3 (accessed 07 April 2021).

Garimella, K. and Tyson, G. (2018), “*WhatsApp doc? A first look at WhatsApp public group data*”, Twelfth International AAAI Conference on Web and Social Media. Twelfth International AAAI Conference on Web and Social Media, available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17865 (accessed 19 October 2021).

Ghasiya, P. and Okamura, K. (2021), “*Investigating COVID-19 news across four nations: a topic modeling and sentiment analysis approach*”, IEEE Access, Vol. 9, pp. 36645-36656, doi: 10.1109/ACCESS.2021.3062875.

Hagberg, A.A., Schult, D.A. and Swart, P.J. (2008), “*Exploring network structure, dynamics, and function using NetworkX*”, in Varoquaux, G., Vaught, T. and Millman, J. (Eds), Proceedings of the 7th Python in Science Conference, Pasadena, CA, pp. 11-15.

Jang, H., Rempel, E., Roth, D., Carenini, G. and Janjua, N.Z. (2021), “*Tracking COVID-19 discourse on twitter in North America: infodemiology study using topic modeling and aspect-based sentiment analysis*”, Journal of Medical Internet Research, Vol. 23 No. 2, p. e25431, doi: 10.2196/25431.

Jiang, M., Gao, Q. and Zhuang, J. (2021), “*Reciprocal spreading and debunking processes of online misinformation: a new rumor spreading–debunking model with a case study*”, Physica A: Statistical Mechanics and Its Applications, Vol. 565, p. 125572, doi: 10.1016/j.physa.2020.125572.

Lewandowsky, S., Cook, J., Ecker, U.K., Lewandowsky, S., Cook, J., Ecker, U.K.H. and Newman, E. (2021), Under the Hood of the Debunking Handbook 2020: A Consensus-Based Handbook of Recommendations for Correcting or Preventing Misinformation.

Lundgren, R.E. and McMakin, A.H. (2013), “Principles of risk communication”, Risk Communication: A Handbook for Communicating Environmental, Safety, and Health Risks. John Wiley & Sons.

Madaan, A., Mittal, A., Mausam, R.G. and Sarawagi, S. (2016), “*Numerical relation extraction with minimal supervision*”, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30 No. 1, available at: https://ojs.aaai.org/index.php/AAAI/article/view/10361 (accessed 17 March 2021).

Martel, C., Pennycook, G. and Rand, D.G. (2020), “*Reliance on emotion promotes belief in fake news*”, Cognitive Research: Principles and Implications, Vol. 5 No. 1, p. 47, doi: 10.1186/s41235-020-00252-3.

McKight, P.E. and Najab, J. (2010), “Kruskal-Wallis test”, The Corsini Encyclopedia of Psychology, American Cancer Society, pp. 1-1, doi: 10.1002/9780470479216.corpsy0491.

Meese, J., Frith, J. and Wilken, R. (2020), “*COVID-19, 5G conspiracies and infrastructural futures*”, Media International Australia, Vol. 177 No. 1, pp. 30-46, doi: 10.1177/1329878X20952165.

Moussaïd, M., Brighton, H. and Gaissmaier, W. (2015), “*The amplification of risk in experimental diffusion chains*”, Proceedings of the National Academy of Sciences, Vol. 112 No. 18, pp. 5631-5636.

Origin of SARS-CoV-2 (2020), “*World health organisation*”, available at: https://www.who.int/publications/i/item/origin-of-sars-cov-2.

Pornpitakpan, C. (2004), "*The persuasiveness of source credibility: a critical review of five decades' evidence*", Journal of Applied Social Psychology, Vol. 34 No. 2, pp. 243-281, doi: 10.1111/j.1559-1816.2004.tb02547.x.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D. (2020), "*Stanza: a Python natural language processing toolkit for many human languages*", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Online, pp. 101-108, doi: 10.18653/v1/2020.acl-demos.14.

Ribeiro, M.H., Gligoric, K. and West, R. (2019), "*Message distortion in information cascades*", The World Wide Web Conference, Association for Computing Machinery (WWW'19, New York, NY), pp. 681-692, doi: 10.1145/3308558.3313531.

Rouhani, S.A., Marsh, R.H., Rimpel, L., Edmond, M.C., Julmisse, M. and Checkett, K.A. (2019), "*Social messaging for global health: lessons from Haiti*", Journal of Global Health, Vol. 9 No. 1, doi: 10.7189/jogh.09.010308.

Singh, M., Jakhar, A.K. and Pandey, S. (2021), "*Sentiment analysis on the impact of coronavirus in social life using the BERT model*", Social Network Analysis and Mining, Vol. 11 No. 1, p. 33, doi: 10.1007/s13278-021-00737-z.

Sun, C., Huang, L. and Qiu, X. (2019), "*Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence*", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). NAACL-HLT 2019, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 380-385, doi: 10.18653/v1/N19-1035.

Tulkens, S. and van Cranenburgh, A. (2020), "*Embarrassingly simple unsupervised aspect extraction*", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, available at: http://arxiv.org/abs/2004.13580 (accessed 15 March 2021).

van der Meer, T.G.L.A. and Jin, Y. (2020), "*Seeking formula for misinformation treatment in public health crises: the effects of corrective information type and source*", Health Communication, Vol. 35 No. 5, pp. 560-575, doi: 10.1080/10410236.2019.1573295.

Vijaykumar, S., Jin, Y., Rogerson, D., Lu, X., Sharma, S., Maughan, A., Fadel, B., de Oliveira Costa, M.S., Pagliari, C. and Morris, D. (2021), "*How shades of truth and age affect responses to COVID-19 (Mis)information: randomized survey experiment among WhatsApp users in UK and Brazil*", Humanities and Social Sciences Communications, Vol. 8 No. 1, pp. 1-12, doi: 10.1057/s41599-021-00752-7.

Walter, N., Cohen, J., Holbert, R.L. and Morag, Y. (2020), "*Fact-checking: a meta-analysis of what works and for whom*", Political Communication, Vol. 37 No. 3, pp. 350-375, doi: 10.1080/10584609.2019.1668894.

Walter, N. and Tukachinsky, R. (2019), "*A meta-analytic examination of the continued influence of misinformation in the face of correction: how powerful is it, why does it happen, and how to stop it?*", Communication Research, Vol. 47 No. 2, pp. 155-177, doi: 10.1177/0093650219854600.

Wardle, C. and Derakhshan, H. (2017), Information Disorder: toward an Interdisciplinary Framework for Research and Policy Making, Council of Europe report, DGI (2017), p. 9.

Woolson, R.F. (2008), "Wilcoxon signed-rank test", Wiley Encyclopedia of Clinical Trials, American Cancer Society, pp. 1-3, doi: 10.1002/9780471462422.eoct979.

Yu, S. and Poger, S. (2019), "Modeling social influence in mobile messaging apps", in Miller, J., Stroulia, E., Lee, K. and Zhang, L.-J. (Eds), Web Services – ICWS 2019, Springer International Publishing (Lecture Notes in Computer Science), Cham, pp. 1-11.

Zeng, J. and Chan, C. (2021), "*A cross-national diagnosis of infodemics: comparing the topical and temporal features of misinformation around COVID-19 in China, India, the US, Germany and France*", Online Information Review, Vol. 45 No. 4, pp. 709-728, doi: 10.1108/OIR-09-2020-0417.

Zeng, R. and Zhu, D. (2019), "*A model and simulation of the emotional contagion of netizens in the process of rumor refutation*", Scientific Reports, Vol. 9 No. 1, p. 14164, doi: 10.1038/s41598-019-50770-4.

Corresponding author

Xingyu Ken Chen can be contacted at: xychen1991@gmail.com