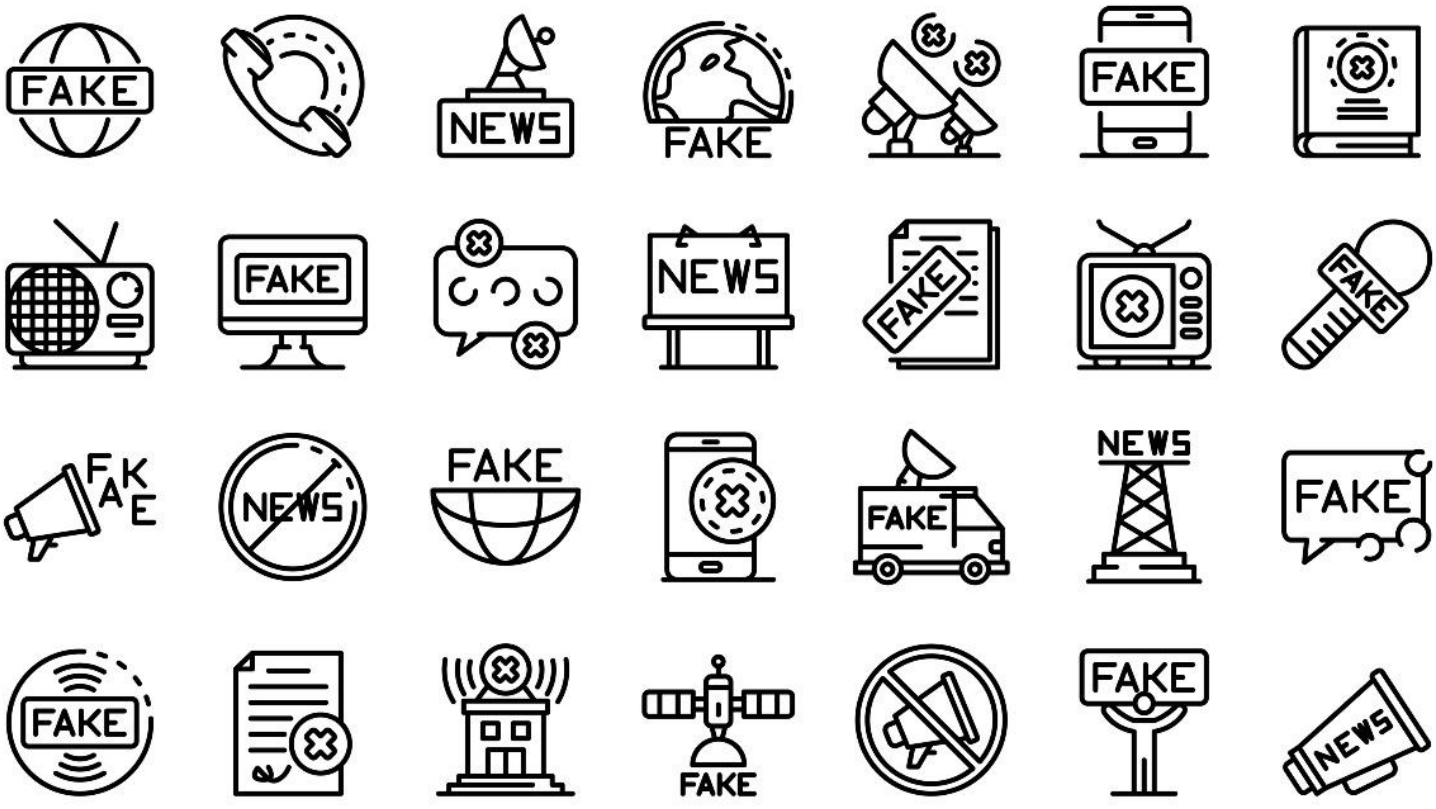
KHAZANAH
RESEARCH
INSTITUTE

WORKING PAPER 03/21 | 23 February 2021

Classification of Information Disorder

Gregory Ho Wai Son and Emir Izat Abdul Rashid



Khazanah Research Institute

The **KRI Working Papers** are a series of research documents by the author(s) containing preliminary findings from ongoing research work. They are a work in progress published to elicit comments and encourage debate. In that respect, readers are encouraged to submit their comments directly to the authors.

The views and opinions expressed are those of the author and may not necessarily represent the official views of KRI. All errors remain the authors' own.

WORKING PAPER 03/21 | 23 FEBRUARY 2021

Classification of Information Disorder

This working paper was prepared by Gregory Ho Wai Son and Emir Izat Abdul Rashid from the Khazanah Research Institute (KRI). The authors are grateful to numerous esteemed persons for providing insightful comments on the paper: Prof. Rosa M. Benito (Universidad Politécnica de Madrid), Harris Zainul (ISIS) and Gayathry Venkiteswaran (University of Nottingham Malaysia). The authors also would like to extend their gratitude to the #NetworkedNation research team, Adam Manaf Mohamed Firouz and Nazihah Muhamad Noor for their valuable comments on the paper.

The authors would like to thank Amos Tong Huai En, Chua Shen Yi, Goh Ming Jun , Lai Kah Chun, Timothy Chan Ying Jie, Wan Amirah Wan Usamah for their diligent assistance in producing this paper during their time as interns at the Institute.

The authors declare no conflict of interest in preparing this Discussion Paper. The paper approached the subject of information disorder via a case study of "Kes-26", a member of KRI's Board of Trustees, Hisham Hamdan who had no involvement nor influence in the preperation of this discussion paper, from initial conception to final publication.

The views and opinions expressed are those of the author and may not necessarily represent the official views of KRI.

All errors remain the authors' own.

Authors' E-Mail Addresses: **gregory.ho@krinstitute.org**, **emirizatrashid@g.ucla.edu**

WORKING PAPER 03/21 | 23 FEBRUARY 2021

Classification of Information Disorder

Gregory Ho Wai Son and Emir Izat Abdul Rashid

Executive Summary

- The press and media agencies played the role of "the fourth pillar of democracy" in disseminating information and facilitating debates in the public sphere. Journalists and reporters hold to a set of standards and responsibilities to uphold truth in facilitating these debates.

- As Malaysia becomes a highly digitalized nation, the prevalence of social media platforms in public discourse have transformed the landscape of opinion formation in the 21st century. Most Malaysians now rely on Facebook and other social media as their main news source.

- Social media platforms have effectively democratized the powers held by reporters and journalists to social media users at large, without the corresponding obligation of holding to high standards of publishing.

- This presents several vulnerabilities to countries and communities. There is evidence to indicate that social media has been weaponized in Myanmar to polarize public opinion, and as a tool to intefere in the US 2016 elections. Without the means to first classify Information Disorder, there is no way to effectively detect or measure these vulnerabilities.

- Evidence from our experiment indicates that human classifications of content were more consistent than human classifications of intent.

- Efforts to combat information disorder may benefit from a combination of human efforts and digitalized solutions. We demonstrate the prospective use of computational methods to complement human intelligence.

- Several policy considerations are discussed: (1) Government—Principles of Objective Fact-checking, (2) Corporations—Civic Responsibilities of handling Information Disorder; and (3) Society—Digital and Online Media Literacy.

1. Introduction

"The limits of my language mean the limits of my world" (Ludwig Wittgenstein)

1.1. Media - The Fourth Pillar of Democracy

The early philosophy of Ludwig Wittgenstein[1] asserts that all philosophical problems are limited by the way in which language convey logical meaning. However, his later philosophy[2] took a turn and posits that the meaning of a word is its use in the language. The latter view highlights the importance of the role of context in determining what people mean. Certainly, with the advent of digital technology, the means of communication in society are very different from Wittgenstein's era. Nonetheless, the arguments raised in both his philosophies remain evermore relevant in today's information age.

Marshall McLuhan argues that it is impossible to understand social and cultural changes without a knowledge of the workings of media[3]. McLuhan argues that media can be categorized in a spectrum of "high definition" to "low definition" media[4]. Books for example, are high definition media as they embody a large volume of information and allow for little interaction on the part of the reciever during the process of transmitting information. Social media on the other hand, are "low definition" in comparison as social media posts contain a smaller volume of information and necessitate the participation of the reciever in a more immersive experience as part of the transmitting process. For "low definition" media, meaning is an emergent process not only from the original content, but also from the various media-reciever interactions (e.g. comments and likes).

On democracy, Amyrta Sen observed that "no substantial famine has ever occurred in any independent and democratic country with a relatively free press."[5] Apart from the three branches of government—"Legislative", "Executive" and "Judicial"—the "Press/Media" is also recognized to be the fourth pillar of democracy. In an ideal democracy, the press plays the role of seeking and disseminating information independently in order to facilitate debates over "truth" and "shared meaning" in the public sphere.

In an increasingly digitalized society, the advent of social media platforms such as Facebook or Twitter has allowed media to be propagated with superior coverage, at unprecedented speeds and at very minimal or no distribution costs. More importantly, there is also a notable transition in the workings of media from high definition to low definition. This implied that the facilitation

[1] Original:Wittgenstein (1921), translation: Wittgenstein and dos Santos (1994)
[2] Original: Wittgenstein (1921), translation: Wittgenstein (2009)
[3] McLuhan and Fiore (1967)
[4] McLuhan and MCLUHAN (1994). Mcluhan uses the terms "hot" and "cold" to refer to high and low definition media respectively.
[5] Sen (1999)

of public debate over "shared meaning" has also transitioned from being dominated by "mainstream press" to a more decentralized system where the "wisdom" (or "madness") of the crowd takes precedence.

For a population size of around 32.7 million people in 2020[6], the Malaysian Communications and Multimedia Commission (MCMC) estimates that there were about 24.6 million users of social networking apps in the country in 2018[7]. Of the total number of social networking users, 97.3% owned a Facebook account, 23.8% owned a Twitter account and 13.3% owned a LinkedIn account.

While the widespread availability of social media most certainly enabled greater freedom of expression, unbounded freedom of expression in the workings of media may not always yield democratic or beneficial outcomes for society. For example, the use of social media has been recognized to play an important role in fuelling violence in Myanmar in September 2017. There were documented accounts of widespread circulation of parallel rumors of imminent attacks which were designed to ignite violence between Muslim and Buddhist communities[8].

Another issue that emerged with the use of social media is the spread of misinformation (and disinformation). Social media platforms employ the use of algorithms whose objective function is to maximize user engagement; a direct consequence of such algorithms is the filtering of media or content for users based on "how much a user interacts with certain 'friends'" and "what type of news feed a user interacts with"[9]. The result is the feeding of problematic media to users that compels real-world actions. For example, the Pizzagate scandal in the US—initially began as a conspiracy theory in *4Chan* (an online platform) —became viral in various social media platforms. This eventually resulted in a firearm incident in the Comet Ping Pong Restaurant on 4th December 2016.

In the era of COVID-19, the perils of misinformation cannot be underestimated. In a joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse and IFRC[10], misinformation has been identified to be harmful—in that "…without the appropriate trust and correct information, diagnostic tests go unused, immunization campaigns (or campaigns to promote effective vaccines) will not meet their targets, and the virus will continue to thrive.". Similarly, disinformation "..is polarizing public debate on topics related to COVID-19; amplifying hate speech; heightening the risk of conflict, violence and human rights violations; and threatening long-terms prospects for advancing democracy, human rights and social cohesion."

Closer to home, the government's initiative to combat misinformation is led by *Sebenarnya.my,* an initiative of the Malaysian Communications and Multimedia Commission (MCMC). *Sebenarnya.my*

[6] DOSM (2020)
[7] Malaysian Communications and Multimedia Commission (2019)
[8] Rio, Victoire (2020)
[9] Bakshy, Messing, and Adamic (2015)
[10] "Managing the COVID-19 Infodemic: Promoting Healthy Behaviours and Mitigating the Harm from Misinformation and Disinformation" (n.d.)

serves as a one stop centre for fact checking potentially errornous media that is flagged to them[11]. In addition to traditional media, the government has also utilized Facebook and Twitter to distribute information.

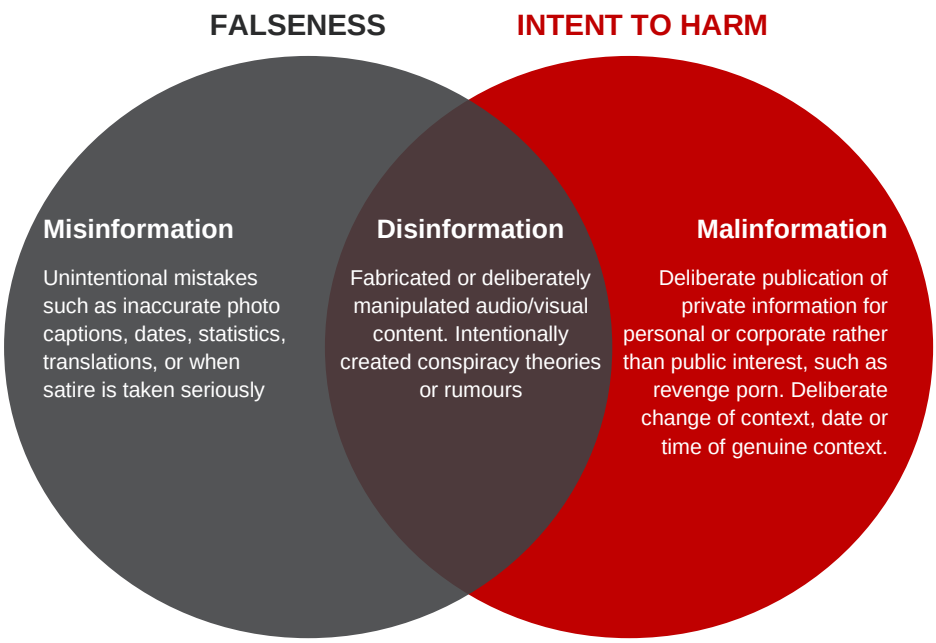
Additionally, a separate survey conducted by Vase.ai[12] highlighted that prior to movement control order (MCO) in March 2020, most Malaysians have relied on Facebook and other social media platforms as their main news source. This is indicative of the prominence and pervasiveness of social media platforms in facilitating public debates in Malaysia today.

1.2. Information Disorder

The widespread use of the term "fake news" has been recognized to be problematic on a few accounts. Firstly, the term has been highly politicized and used as more than just a label for false and misleading information. Fake news has been deployed as a weapon against news agencies as a way of undermining reporting that people in power do not like[13]. Secondly, the term also conflates the various definitions of Information Disorder. These definitions were constructed to quantify and discuss a phenomenon that is more complex than what "fake news" is able envelop[14].

Figure 1 defines Information Disorder as a Venn Diagram along the dimensions of "Falseness" and "Intent to Harm".

Figure 1: Types of Information Disorder



Source: Adapted from Wardle and Derakhshan (2017)

[11] Authors' edit to manuscript: a previous version of this paper misstated the role of the MCMC in other government initiatives.
[12] Vase.ai (2019)
[13] Ireton and Posetti (2018)
[14] Ibid

"Falseness" simply represents the presence of content that is untrue. "Intent to Harm" represents the presence of content that either is specifically (or maliciously) constructed to bring physical harm, or is an "attack on dignity" to harm the reputation of a person, institution or social group. Along these two dimensions, Information Disorder is then categorized as Mis-Information, Dis-Information or Mal-Information:

Table 1: Three Categories of Information Disorder

Category	Description
Mis-information	False information is present, but no harm is intended.
Dis-information	False information is knowingly shared to cause harm.
Malinformation	Genuine information is shared to cause harm, often by moving private information to the public sphere.

Source: Adapted from Wardle and Derakhshan (2017)

However, the use of such terminology can sometimes be confusing or inconsistent in experimental or analytical settings[15]. This is because the key to effectively employing the definitions lie in the ability to confidently label content along two dimensions. Firstly, to label some media content as containing "falseness", the encoder must claim **access to truth**. Secondly, to label some media content as containing "intent to harm", the encoder must have the correct interpretation of **authorial intent**. Where there is reasonable uncertainty with regards to these two dimensions, it can be difficult to make a case for the presence of Information Disorder, or which category of Information Disorder is present in some media.

Moreover, these difficulties necessarily extend beyond the experimental and analytical settings. In a legal setting, the "burden of proof" falls on the part of the plaintiff who has to make a case for which the defendant is guilty of either "falseness" or "intent to harm". Additionally, the plaintiff has to defend his position beyond reasonable doubt ("standard of proof") in the case of criminal proceedings. Whatever judgements that arise from such cases also form a precedent for future cases.

On 11th April 2018, the Anti-Fake News Act 2018 (Act 803) was passed by the Malaysian Parliament[16]. In an interview with the minister in the Prime Minister's Department responsible over the bill, Azalina Othman explained that the act was passed to deal with "…the issue of dissemination of fake news is a global problem, following the technological communication revolution, which is happening at a rapid pace. Of late, Malaysia has faced numerous challenges as an effect from fake news which not only confuses the public but can also threaten the safety, economy, prosperity and well-being of the people and the country"[17].

This act has since been repealed on 9th October 2019. Explanatory statements accompanying the bill to repeal Anti-Fake News Act 2018 (Act 803) states that Act 803 is no longer relevant given

[15] Wu et al. (2019)
[16] Malaysian Federal Gazette website:Link. Extracted: October 2020
[17] Ngah (2018)

changes in the Penal Code (Act 574), the Printing Presses and Publications Act 1984 (Act 301) and the Communications and Multimedia Act 1998 (Act 588)[18].

Wu et al. (2019) argues that in the spectrum of media containing Information Disorder, distinguishing between "Mis-information" and "Dis-information" is relatively difficult. The key difficulty lies in determining whether the content was intentionalley and deliberately constructed to deceive, to mislead, or to cause harm. Wu et al. (2019) organizes Information Disorder as follows in Table 2.

Table 2: Other categorizations of Information Disorder

Category	Description
Unintentionally spread misinformation	Instead of wanting to deceive, a user tries to inform their social network about a certain issue or situation
Intentionally spread misinformation	Usually writers and coordinated groups of spreaders who have a clear goal and agenda to compile and promote misinformation
Urban legends	Intentionally spread misinformation on fictional stories. Can often be for entertainment
Fake news	Intentionally spread misinformation that is in the format of news. *(original definition before the popularization of the term)*
Rumors	Unverified information (can be true)
Crowdturfing	Inflation of support (likes) via the use of marketing agents / bots
Spam	Unsolicited information that unfairly overwhelms recipients
Troll	Cause disruption and arguments in a discussion
Hate speech	Content that targets certain groups of people, inciting hatred and violence

Source: Adapted from Wu et al. (2019)

Conversely, harm can be also be unintentionally but wrongfully inflicted on individuals, institutions, or social groups in the pursuit of truth. Should these content be categorized as containing Information Disorder?

2. Objectives of Research

This section provides an outline of the research objectives of this paper. Firstly, the paper wishes to investigate the subtleties and problems associated with categorizing Information Disorder in the context of human assessments. Given the issues highlighted above, the paper investigates whether human assessments are able to consistently detect the presence of Information Disorder and consistently label them according to the various categorizations of Information Disorder.

Secondly, the paper wishes to investigate if computational approaches can be employed to aid with the classification of Information Disorder. We investigate if unsupervised learning, specifically whether Latent Dirichlet Allocation (LDA) can be implemented to augment human assessments at a larger scale.

[18] Link to the bill.

Given that the scale, openness and timeliness of social media have largely transformed the role of media as the fourth pillar of democracy, the paper examines if problems that emerge from an increasingly digitalized society can also be battled using digitalized solutions that emerge from data science and machine learning.

We aim to do this by studying tweets regarding a specific COVID-19 patient that was the subject of widespread public discussion in Malaysia in early 2020.

3. Methods and Results

3.1. Context of the case study

At the time of data collection in early 2020, we only had access to a standard Twitter developer account. The standard account has limitations with regards to the breadth, depth and timeframe of search and extraction of tweets. With these limitations in mind, this paper approaches Information Disorder in Malaysia by use of a case study of "Kes-26". "Kes-26" was the 26th person in Malaysia to have tested positive for COVID-19, and was initially identified by KKM[19] to be the index case of over 20 other infections.

At the onset of COVID-19 in pre-lockdown Malaysia, "Kes-26" was interesting because its prominence in social media led to speculations on the identity of the patient especially following the circulation of a rumor of his attendance at a particular political event leading up to his diagnosis[20]. Following these events, there were various forms of Information Disorder being spread on social media concerning him. On 6th March 2020, "Kes-26" released a public statement, sharing the facts and clarifying the statements that were being made about him[21].

The case study is based on a relatively small corpus of Twitter data. We generated the dataset by collecting tweets surrounding "Kes-26". For data collection, we developed a crawler based on Twitter's API, which allows for a filtered collection of real-time tweets based on pre-specified keywords. The keywords which were employed are summarized in Table 3 below.

Table 3: Keywords used for filtered collection of tweets

Keywords	Description
"Kes-26", "Kes 26", "Case-26", "Case 26"	Keywords directly refering to Kes-26
"Hisham Hamdan", "UDA Chair"	The "identities" of Kes-26

Based on the filtered collection, we extracted a total of 2,015 tweets, posted by 1,569 unique twitter handles, over the period of 28 February 2020 to 10 March 2020.

[19] Link : KKM Portal MyHealth on Twitter
[20] Sarawak Report (2020)
[21] Appendix B: Kes-26 Public Letter

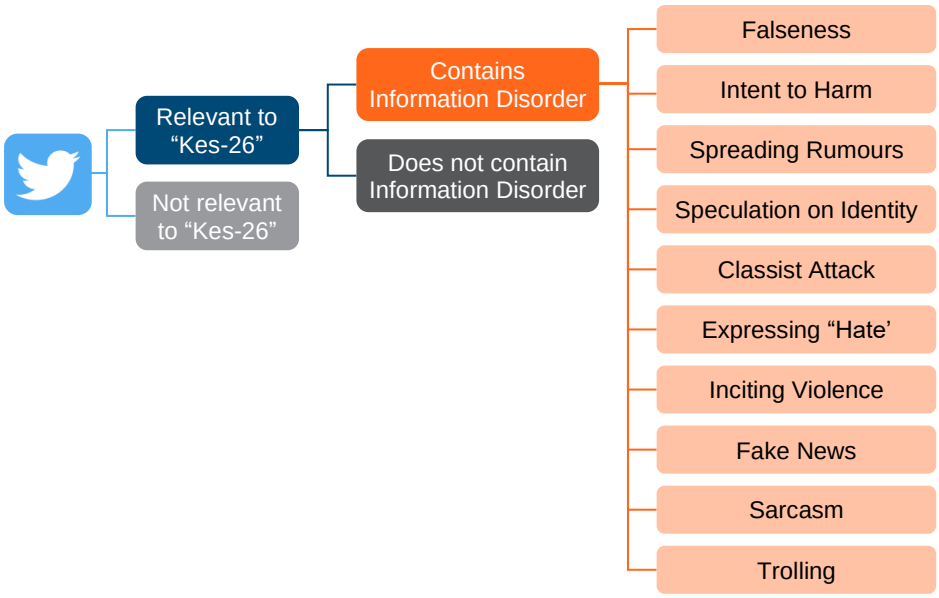
3.2. Human assessments

The 2,015 tweets were distributed among 7 coders in a way where each tweet was classified by a minimum of 3 coders. Each coder classified 860+ tweets per person. To ensure that the classification was done in a consistent and objective way, coders were:

1. Exposed to the events surrounding Kes-26
2. Exposed to the literature on the categorization of Information Disorder
3. Given a randomized set of 100 tweets as a practice set before the actual classification exercise. Post-practice, two meetings were held to discuss issues and iron out interpretive dissonance on what each class represents.

Figure 2 below represents the classification of Information Disorder that we employ in this exercise.

Figure 2: Classification of Information Disorder



Based on Figure 2, coders first classified if a particular tweet is "relevant" or "not relevant" to "Kes-26". As we are only interested to study Information Disorder within the context of our case study, "non-relevant" tweets were filtered out in order to generate a dataset that is specific to "Kes-26". Some of the non-relevant tweets include tweets promoting mobile phone cases, tweets reporting COVID-19 cases in other countries, or even legal cases in other countries.

For relevant tweets, coders then determined if the tweet contains Information Disorder, or if it does not contain Information Disorder. If the tweet contains Information Disorder, coders had the additional task of identifying if the tweet contains particular definitions of Information Disorder based on our literature review[22]. These definitions are not mutually exclusive (the presence of one definition does not necessitate the absence of another). The authors also made

[22] These definitions were developed mainly from Wardle and Derakhshan (2017) and Wu et al. (2019).

the decision to add the following definitions in Table 4, based on feedback from the practice exercise:

Table 4: Additional Classifications of Information Disorder

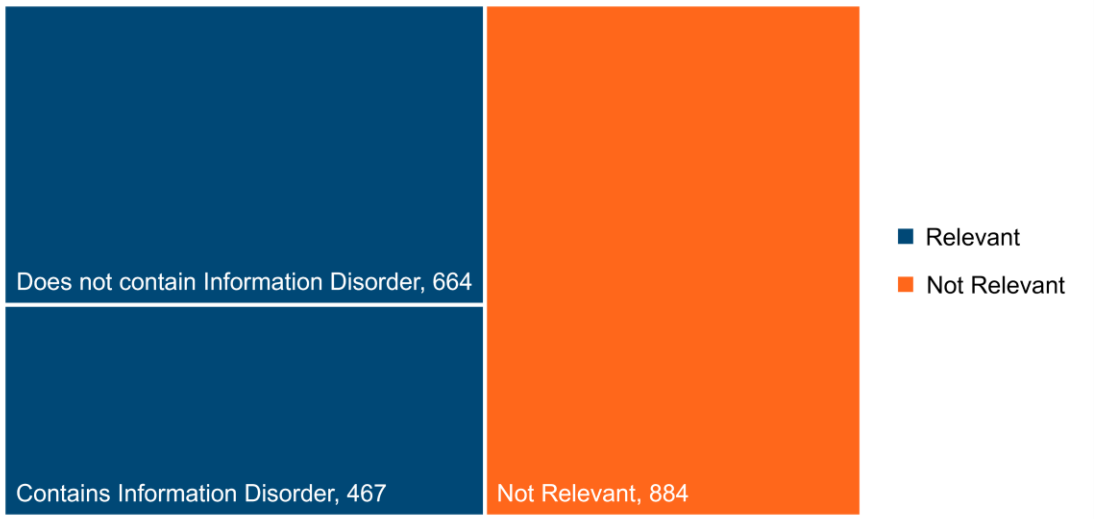
Category	Description
Speculation on identity	Intentionally trying to reveal the identity of a COVID patient
Classist attack	Expressing prejudice against a particular group based on class
Expressing hate	Use of pejorative and discriminatory language with reference to a person or group based on who they are[23].
Inciting violence	Advocating a crime, or injury to person or social group

Moreover, these additional definitions may have more serious implications than the earlier definitions. For example, the spreading of rumors can be combated by sharing facts and the truth to dispel misinformation. However, the strategy to combat content with pejorative language or content even suggesting physical harm to a person or a social group may require more than just the sharing of facts. The law of Malaysia makes provisions for this in Act 574 (Penal Code) Section 153[24].

Descriptive Results

This section summarizes the descriptive results that emerge from the human assessments. Figure 3 below describes the proportion of tweets that contain Information Disorder in the set of tweets relevant to our case study:

Figure 3: Proportion of Tweets with Information Disorder



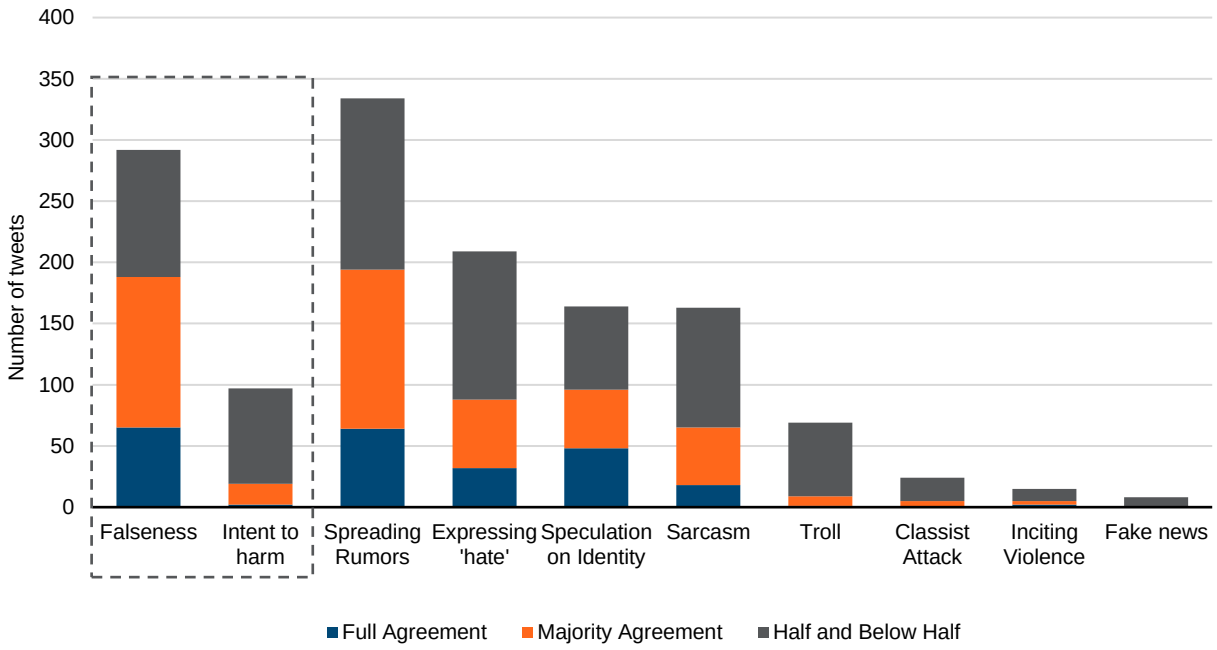
Source: Authors' own calculation

[23] Guterres (2019)
[24] Attorney General's Chambers of Malaysia, Source: document.

Out of the total 2,015 tweets that were extracted, 884 were considered to be not relevant to our case study[25]. Of the remaining 1,131 tweets relevant to the case study, only 467 (41.3%) were considered to contain some form of Information Disorder. However, the quantification of what proportion of tweets contain Information Disorder would also vary across country, topic and time periods. Another study that employed human assessments over a much larger volume of tweets has found about 23.46% of tweets to be perceived as not credible[26].

For the 470 tweets that contained some form of Information Disorder, Figure 4 describes the consistency of human assessments[27].

Figure 4: Consistency of Human Assessments



Source: Authors' own calculation

Observations

Firstly, over the period of 28 February 2020 to 10 March 2020—of the tweets that were relevant to "Kes-26"—most tweets (664 [58.7%] out of 1,131) did not contain Information Disorder.

Secondly, coders were able to consistently differentiate tweets which contain Information Disorder and tweets which did not contain Information Disorder[28].

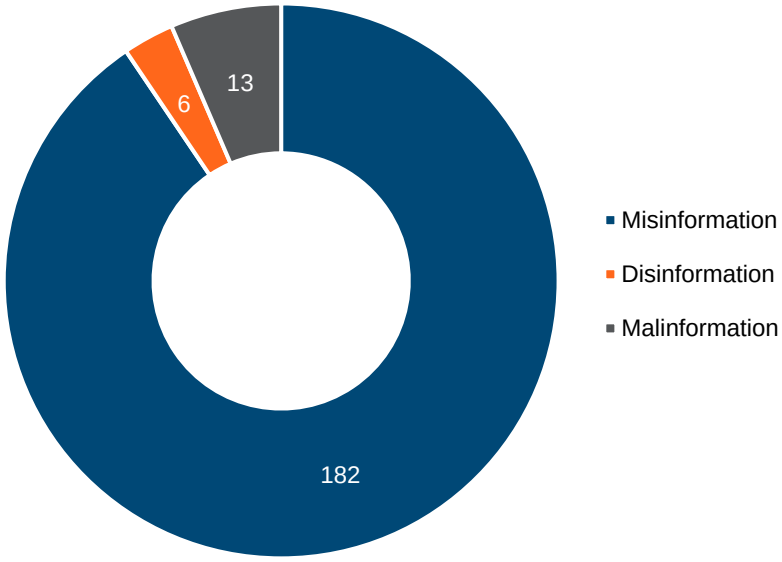
[25] We only consider definitions to be "present" for tweets where there is full agreement or majority (2/3 or 3/4) agreement.
[26] Mitra and Gilbert (2015)
[27] In considering Information Disorder, we employed a more stringent measure. A tweet is considered to contain some form of Information Disorder if there is at least one human that identifies the presence of a particular definition of Information Disorder.
[28] Majority and Full Agreement is observed to be at 86.4%.

Thirdly, certain dimensions of Information Disorder were classified relatively more consistently than others. The dimensions of "falseness", "spreading rumors", "expressing hate" and "speculation on identity" all had a sizeable amount of full and majority agreement. However, when it comes to the presence of "intent to harm", "trolling", "classist attacks", "inciting violence" and "fake news", the classifications the coders decided on did not agree with one another. In fact, there were close to zero cases of full agreements over the aforementioned dimensions.

The presence of such ambiguity in the classification of these dimensions could be a signal of the difficulties people face in reading intent, and subsequently making a judgement based on that.

Of the 470 tweets containing Information Disorder, the presence of "falseness" far outweighs "intent to harm". For the tweets where "falseness" and (or) "intent to harm" are recognized to be present at majority and full agreement, Figure 5 describes the distribution of Information Disorder as defined by [29] in our human assessments.

Figure 5: Information Disorder in the Case Study of "Kes-26"



Source: Authors' own calculation

Potentially, this implies that the Information Disorder tweets were mostly misinformation—tweets containing content that have an element of falseness, but was not constructed maliciously with intent to harm. Unfortunately, we are unable to determine if this is a feature of the case study, or if it is a result of ambiguity in the classification of "intent to harm", as outlined in the next section.

[29] Wardle and Derakhshan (2017)

Summary of Reflection Essays by Coders

Coders involved in the human assessment were also asked to write a reflective essay to summarize their experience in this classification exercise. This section seeks to summarize key reflections emerging from the exercise.

Firstly, all coders experienced difficulties with regards to **assessing intent and motive**. It was argued that to be able to objectively determine if a person "intends to harm" another, the motive of writing the particular tweet has to be made known. In some cases, coders were able to rely on verbal cues, such as the CAPTITALIZATION OF ALL LETTERS or the use of strong emotional words. However, to truly know a person's motivations is almost impossible: the practice of classifying content based on the **authorial intent** can be somewhat subjective.

Secondly, even when one had access to all the facts, there is still ambiguity that arises from the use of language. In social media platforms like Facebook or Twitter, what is posted is often unscripted, unfiltered sentiment of the populus. In many cases, what has been said can be very different from what one intends to say. Moreover, in a world where truth can often be relative, should people be labeled as spreading misinformation based on an incorrect or inaccurate use of words that both they themselves and the rest of the world do not fully understand?

Thirdly, Malaysia is a multilingual country—the result of a melting pot of different ethnic groups and cultures. Many tweets were written in a combination of at least two or more languages with use of shorthands, abbreviations and memes. In many cases, coders had to understand how these features were used in order to gauge the intent behind the tweets.

3.3. Natural Language Processing (NLP)

Up to this point, the human assessments have emphasized the importance of understanding context, the structure and rules of language, and having the ability to correctly identify motive and intent in order to objectively identify the presence of Information Disorder.

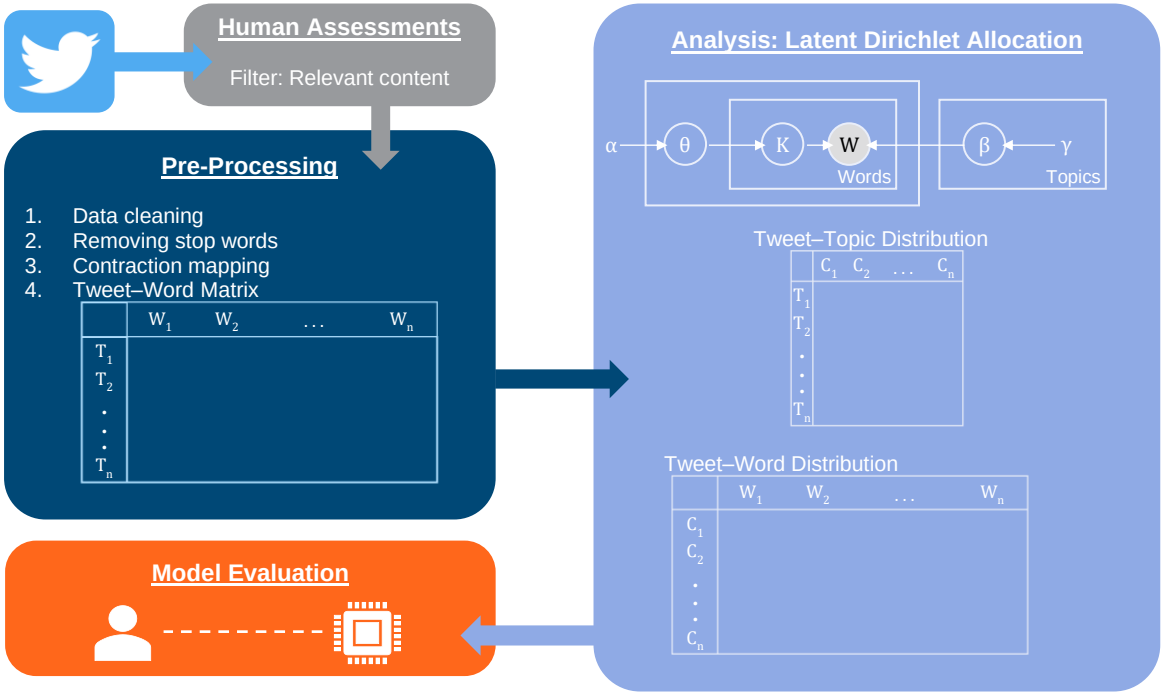
However, in a digitalized society where millions of new tweets are generated every single day, the scale and time-sensitivity of having humans assess large volumes of tweets would be severely impractical and expensive. A digitalized society requires digitalized solutions or augmentations that can greatly ease the workload of classifying Information Disorder.

In this section we explore the use of NLP techniques as a potential enabler to augment human assessments. NLP represents a body of statistical tools, techniques and algorithms used to process natural language based data (normally unstructured) like text, documents or speeches.

Data Processing & Analytical Pipeline

A visual description of our data processing and analytical pipeline is illustrated in Figure 6 below:

Figure 6: Data Processing and Analytical Pipeline



Based on Table 3, we have extracted a total of 2,015 tweets, posted by 1,569 unique twitter handles, over the period of 28 February 2020 to 10 March 2020. The tweets were filtered to obtain a dataset that only contained tweets that were relevant to our case study[30]. This dataset is henceforth referred to as the "Kes-26" corpus. The corpus is then subjected to standard NLP data pre-processing techniques. Table 5 describes the various pre-processing steps that were conducted preceding the analytical steps.

Table 5: NLP Data Pre-processing Steps

Step	Processes	Description
1. Data cleaning	• Removing html tags, Twitter handles and special characters • Correcting spelling errors	These elements have to be addressed as they contribute to more noise in the dataset.
2. Handling "stop words"	• Removing stop words	Stopwords ("a", "and", "the") are words that appear very frequently, but have little significance analytically.
3. Contraction mapping	• Map contractions, abbreviations and shorthands	Contractions ("you're", "you've") are shortened words or syllables. These elements are mapped to ensure consistency in the use of words.

[30] Relevent tweets were defined based on human assessments.

After the pre-processing stage, the corpus is then tokenized[31] to generate a traditional document–term matrix (which is henceforth defined as the tweet–word matrix [TW matrix]). Elements of the TW matrix are computed according to the frequency of appearance for each word in each tweet. In the traditional NLP literature, the TW matrix is classified as a "Bag of Words" model. One consequence in employing "Bag of Words" models, is that these models dissolve any information in semantics, structure, sequence and context when coarse-grained[32]. As described in Figure 6, the most basic form of the TW matrix is simply a frequency count of the occurrence of each word, in each tweet.

Descriptive Results

Before analyzing the data further, we first used information retrieval methods, namely "Term Frequency–Inverse Document Frequency" (TF–IDF) and correlation analysis to extract a general overview of the Kes-26 corpus.

The frequency of a word seems intuitive at first—the more a word is repeated, the more we know what kind of words are being used; however, word frequencies are often unhelpful to capture the words that could make a sentence meaningful. For example, in our data set, the word "kes", "26", "hisham", "covid", and "case" are in the top 20 words used, by virtue of these words being used as parameters for the filtered collection of tweets. However, those words do not provide any meaningful information regarding the tweet data sets as they were employed on Twitter's API specifically to retrieve tweets which contain these words.

To generate a general description, we therefore employed TF–IDF instead of word frequency to gauge the Kes-26 corpus. TF–IDF assigns an index value to every word in the corpus based on the following inputs:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad \ldots(1)$$

$$IDF(w) = \log\left(\frac{N}{df_i}\right) \quad \ldots(2)$$

$$TF - IDF = \frac{n_{i,j}}{\sum_k n_{i,j}} \; x \log\left(\frac{N}{df_i}\right) \quad \ldots(3)$$

A TF–IDF[33] value increases proportionally with the number of times a particular word is repeated in a single tweet, but is offsett by the number of times the word is being used in other tweets. For example, common words that are being used many times in one tweet, but also at the same time being used across the tweet data set, will most likely be lower than words that are being used less in a single tweet but less prevalent across the data set. An ideally high TF–IDF score will be words that are used a lot in a single tweet but is not repeated again in the data set. Figure 7 below

[31] Breaking sentences into linguistic units – case words

[32] Sarkar (2016)

[33] The algorithmic form of this equation applies $\log\left(\frac{N}{df+1}\right)$ to avoid division over 0.

describes the top 10 highest valued words based on TF–IDF[34] while Figure 8 describes the top 10 highest word frequencies of the corpus.

Figure 7: Top 10 TF-IDF Valued words in the corpus

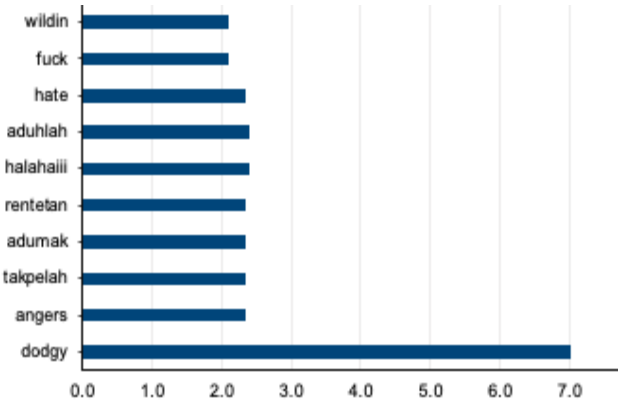
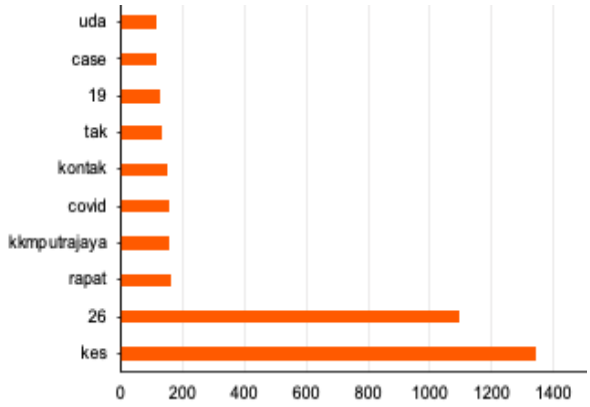


Figure 8: Top 10 words by frequency in the corpus



High TF–IDF words paint a better picture for the overall sentiment of the tweet data set than word frequencies. For example, the words "tak", "kontak", "rapat" reflect common features that provide little contextual information. On the other hand, words like *the f-word*, "dodgy", "adumak" or "halahaii" reflect stronger emotional responses that relate to the context of the corpus.

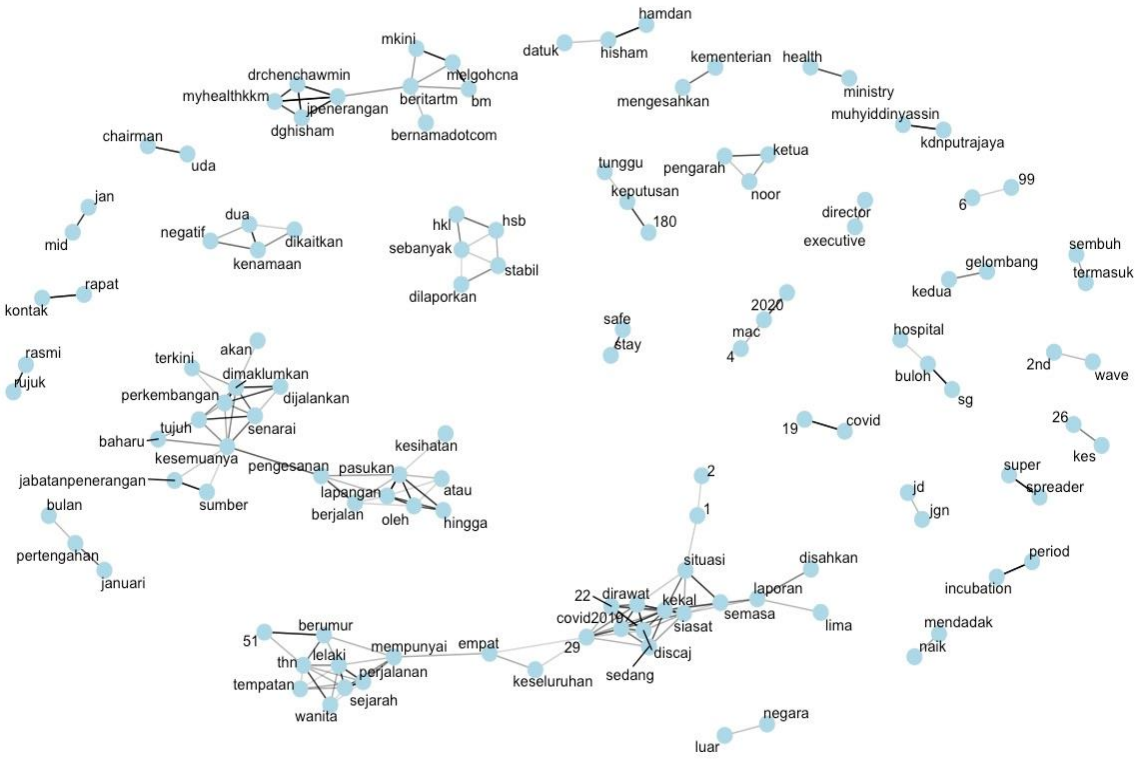
Next, to infer trends in the underlying topics and ideas that were discussed in the corpus, the TW matrix is reorganized according to bi-gram[35] pairs of words (instead of individual words, bigram tokenization breaks the sentence into word pairs). Then the correlation coefficient ρ is computed for each word pair. Using ρ, we generated a correlation network of words (Figure 9) "naturally" occuring in the corpus.

Figure 9 describes the word map based on the frequency of words appearing in succession. As observed in the correlation network, words that frequently appear together agglomerate closer together to form "islands". "Islands" represent themes, entities, events and other features of the corpus.

[34] Malaysia is a melting pot of many cultures, religions and languages. For every non-English word that is used in our analysis, a corresponding definition may be found in Appendix A

[35] A bi-gram represents a sequence of two adjacent words appearing in each tweet.

Figure 9: Correlation network between words (ρ > 0.5)



The two major word "islands" are composed of official Malay words, which we conclude are announcements or reports from news agencies and the Ministry of Health. On the other hand, the word "super" and "spreader"—a recurring allegation—appears to form their own word island. Other notable word islands represent the identities associated with "Kes-26", places which he has visited, the characteristics of COVID-19, and other related themes.

The generation of this correlation network makes no prior assumptions on the themes save for the "filtered words" applied in the Twitter's API. This indicates that the method may be employed as a precursory diagnostic tool that may help in human assessments, provided that the "filtered words" were selected accurately.

Topic Analysis: LDA

For human assessments of Information Disorder, two ingredients are necessary—**access to truth** and the **ability to correctly diagnose authorial intent**. In this section, we explore the possibility of classifying Information Disorder without making any apriori assumptions on the characteristics of Information Disorder.

To do this, we employ the use of unsupervised classification, a set of techniques in machine learning whose goal is to correctly group objects based on shared "similarity(-ies)" over a set of dimensions. In particular, we employ the use of a topic modeling method called Latent Dirichlet[36]

[36] The Dirichlet distribution is generalized from the Beta distribution for multiple random variables.

Allocation (LDA) as first described in Blei, Ng, and Jordan (2003) in the context of machine learning.

Fundamentally, LDA is a generative probabilistic graphical model based on a three-level hierarchical Bayes model, in which each "tweet" of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is in turn, modelled as an infinite mixture over an underlying set of topic probabilities. The basic idea is that "tweets" are represented as random mixtures of "latent"[37] topics, and topics are characterized by a distribution of words[38].

In standard plate notation, Figure 10 below describes the representation of LDA as a Probabilistic Graphical Model (PGM).

Figure 10: Plate notation of LDA

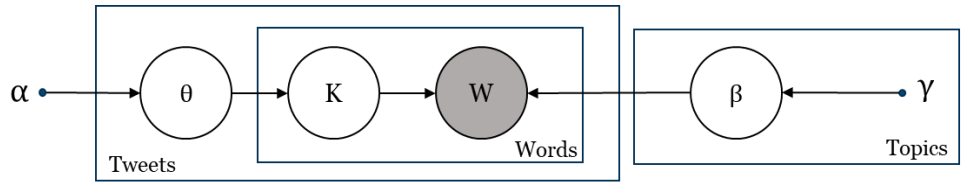


Plate notation is a common method of representing variables which repeat in a graphical model. A plate (the boxes in Figure 10) represents repeating sub-graphs. The variable names are defined as follows:

- W_{ij}: A particular word
- K_{ij}: The topic for the j-th word in tweet i.
- θ_i: Topic distibution for document i
- β: Dirichlet prior on the per-topic word distribution
- α: Dirichlet prior on the per-tweet topic distribution
- γ: parameter vector for each topic in β.

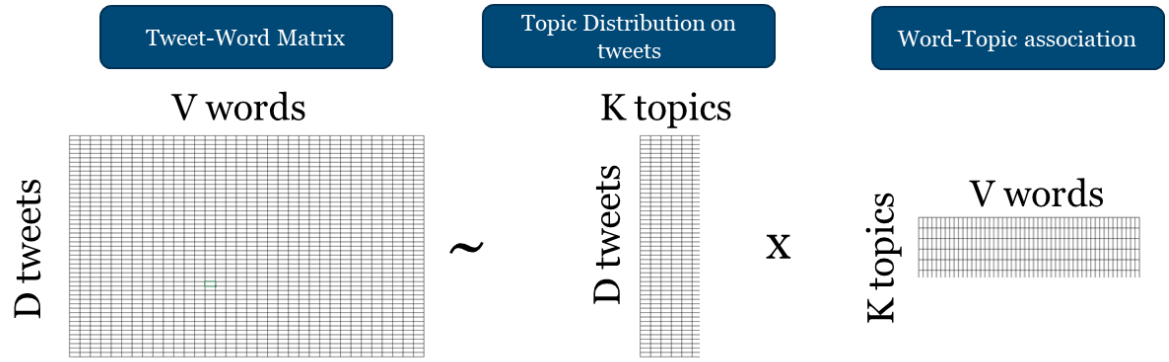
The "word" node is greyed because it represents the only variable in LDA that is directly observed, all other variables are "latent" in the construct of the model.

- A higher α value results in tweets being composed by just a few topics, while a lower α value results in tweets being composed by all topics.
- A higher β value results in topics containing a mixture of most of the words in the corpus, while a lower β value results in topics containing a mixture of few of the words in the corpus.

Through optimization, LDA discoveres themes through posterior inference[39]. The only input required from the analyst is the total number of topics that is contained in the corpus. A simplified description of what we are doing is described in Figure 11 below:

[37] The term latent may be used interchangeably with the term "hidden" in statistical learning.
[38] Blei, Ng, and Jordan (2003)
[39] Blei, Ng, and Jordan (2003)

Figure 11: Depiction of LDA in Matrix Form



The mathematical formulation to statistically infer the Topic–Word and Tweet–Topic Distribution as follows:

$$P(\theta, K, \beta \mid Corpus; \alpha, \gamma)$$

Specifically, the equation "learns" the joint posterior probability of θ and β given the tweets in our corpus, with the parameters α and γ, representing tweet-topic and topic-word distributions respectively.

One potential issue with regards to LDA is that the algorithm requires the analyst to first specify the number of topics, K. K can be specified either based on field expertise which in our case study suggests any number between 2 and 11. The number 2 being simply "Information Disorder" vs. "no Information Disorder", while 11 represents the various definitions in Figure 4 including "no Information Disorder". Alternatively, the number of topics, K can be determined intrinsically based on K-fold Cross-Validation[40], optimizing for 2 common measurements in Topic Model Analysis as described in Table 6 below.

[40] K-Fold Cross Validation is a statistical method to estimate prediction error. The method splits the data set into K roughly equal sized parts, in which the model is trained on (K-1) parts. Prediction error is then estimated from the validation set (K^{th} part). K=5 in our analysis.

Table 6: Perplexity and Coherence

Measure	Functional Form	Description
Perplexity	$= \exp\{-\frac{\sum_{d=1}^{M} \log P\ (w_d \mid \emptyset, \alpha)}{\sum_{d=1}^{M} N_d}\}$	Perplexity represents the probability of the model, predicting the validation set.
Coherence	$= \sum_{i<j} Score_{sim}(w_i, w_j)$ where: $Score_{sim} = \log\frac{D(w_i, w_j) + 1}{D(w_i)}\}$	Coherence measures to what degree topics exhibit semantic similarity.

Figure 12 and Figure 13 below describe the two measures varied over a different candidate number of topics.

Figure 12: Model Perplexity

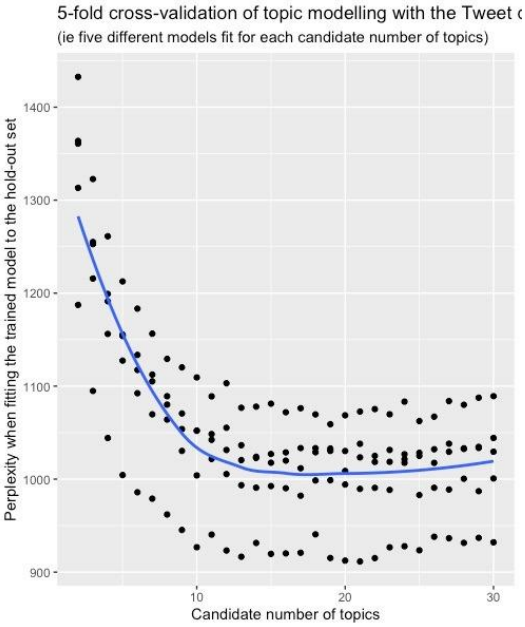
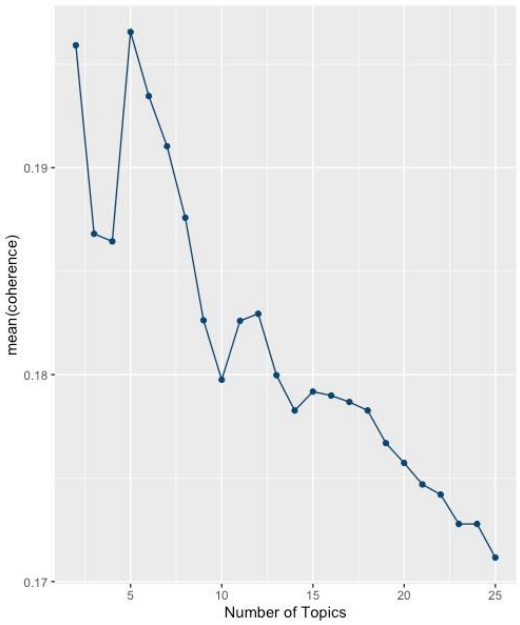


Figure 13: Mean Coherence score



By performing a 5-fold cross-validation on the dataset, model perplexity converges at slightly above 1,000 above 15 candidate number of topics. This measure suggests that a model with above 15 candidate number of topics would better fit the validation set as compared to models with candidate numbers fewer than 15.

However, while perplexity is useful for evaluating a predictive model, the features generated do not address the more exploratory goals of topic modelling[41]. More importantly, perplexity does

[41] Chang et al. (2009)

not help explain the semantic structure between topics, as it is simply a log-likelihood measurement of what words predict the held-out model (validation set).

For that, topic coherence is used to measure the degree of semantic similarity between topics. A higher coherence score means that the topics are more distinguishable from each other. An ideal candidate number of topic choice will have the highest mean coherence of all topics across all topics. We analyzed the mean coherence value as low as 2 topics to 25 topics.

As observed in Figure 13, mean coherence peaks at 5 topics and shows a steady declining trend moving forward. The measure of coherence suggests that there are inherently 5 latent topics in the "Kes-26" corpus that give the best result in observing inter-topic distinguishability and intra-topic similarity.

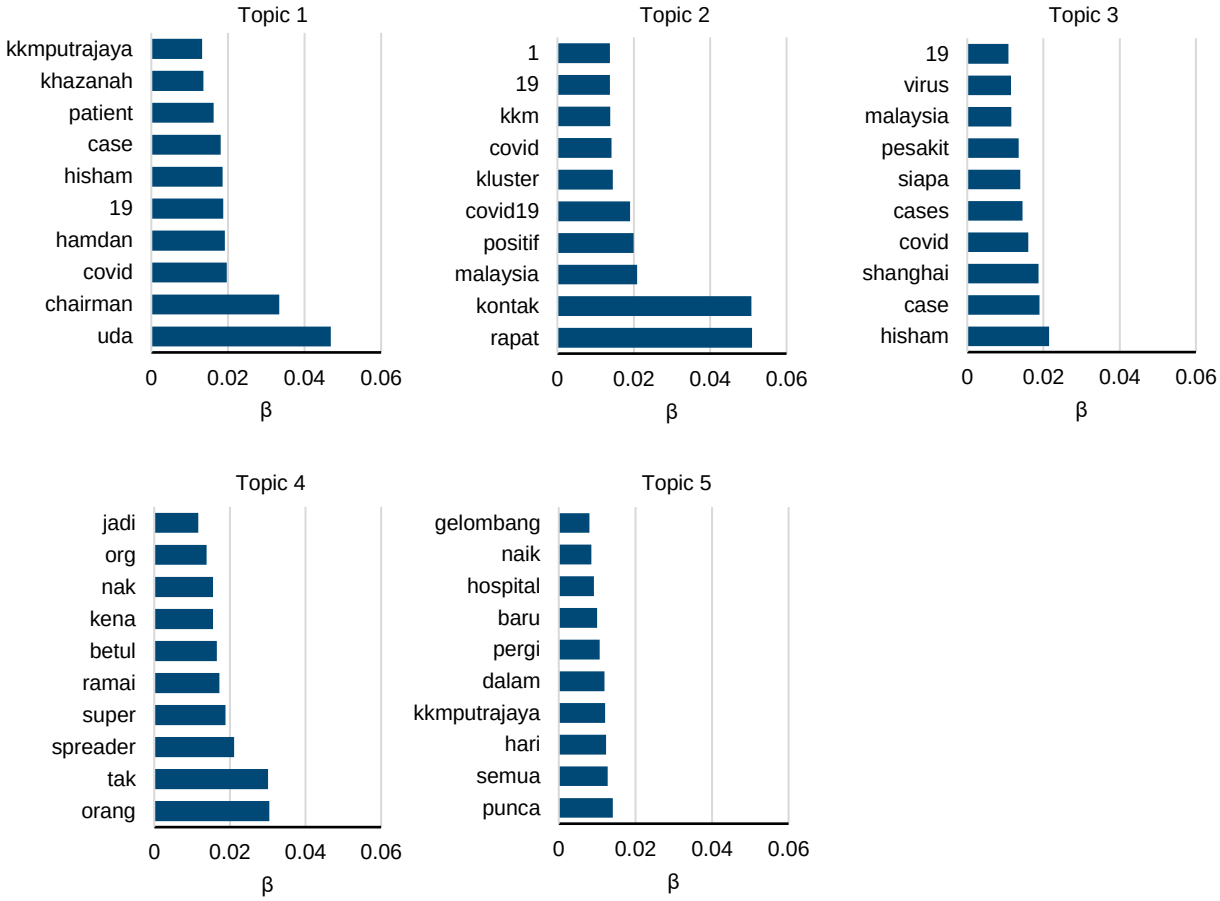
However, there is also a similar peak at 2 topics. In the case of 2 topics, the model seems to categorize tweets according to those tweeted by media and news agencies, and those tweeted by other individual users on Twitter.

LDA: Topic–Word Association

Within the framework of LDA, every word is associated with topics based on β (Figure 10). As a result, each topic can then be described as a distribution of words[42]. By rank-ordering the β-values for each topic, we can then explore the semantic structure of what the topic represents. For each of the 5 topics, LDA assigns β-values for a total of 1,828 words. Figure 14 describes the top 10 words by each topic.

[42] Technically, a K-dimensional discrete representation.

Figure 14: Top 10 terms for each topic group according to their assigned β-values



While top terms are a good start to gauge the composition of the distinct topic groups, the sheer scale in the number of words can be informationally overwhelming. The use of TF–IDF can further simplify the process by singling out words that are more polarized in the distribution. High polarization indicates clear distinctions between groups while high frequency words with larger spread distribution indicates that the word is being used across the entire data set. For example, Figure 15 and Figure 16 describes the β-value of 5 highest TF–IDF words and 5 highest frequency words respectively against the topic number.

Figure 15: High TF–IDF Topic–Word distribution

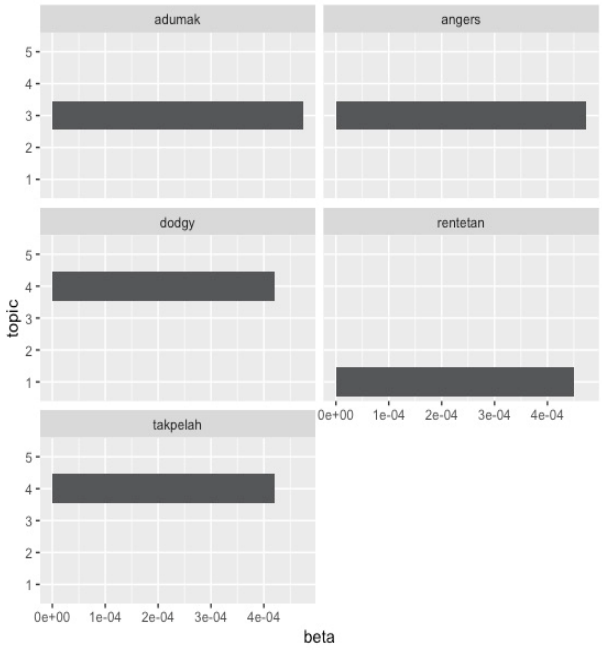
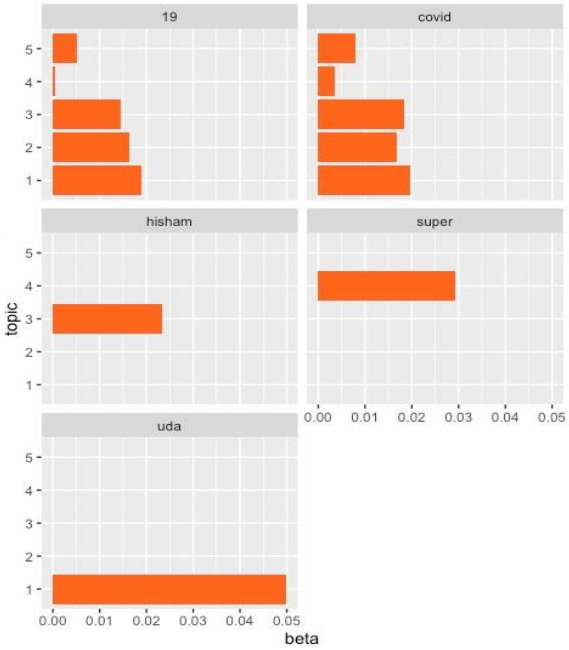


Figure 16: High Frequency Topic–Word Distribution



Firstly, high TF–IDF words are observed to be more polarized than high frequency words. This is expected because words that appear less frequently in the corpus give more informational value as compared to words that appear in most tweets. Hence, specific words can be associated more closely to the individual tweet. The word "dodgy" and "takpelah" both have similar distribution (highly probable to be classified in topic 4), while "adumak" and "angers" (highly probable to be classified in topic 3). Interestingly, none of the high TF–IDF words belong in topic 2, a strong sign that topic 2 is less likely to have "outlier" words.

High frequency words on the other hand are less polarized. For "covid" and "19", there is a high spread. This shows that the usage of those two words are common throughout the data set. By the same token, the word "covid" and "19" have a similar β-value trend except for a slight bump in topic 4 for "covid". These two similar distributions indicate a strong keyword relation – as we know, "covid" and "19" appears together a lot, but "covid" can also be used without "19", hence the small difference in topic 4.

It is observed that the word "super" is dominated by topic 4, alongside "dodgy" and "takpelah", while "hisham" is overwhelmingly in topic 3, shared by "adumak" and "angers". Similar probabilty distributions indicate that these words share some form of "commonality" that a human otherwise could not be able to decipher.

The word–topic distribution shows the remarkable ability for the algorithm to detect similarities between seemingly unrelated words. The word–topic distribution does not suggest that some words are intrinsically related with Information Disorder. It shows that tweets containing Information Disorder share word similarities, as indicated by highly polarized word–topic distribution and strong keyword relations between other words.

Based on the results, we can establish some word patterns using a word–topic probabilty distribution; the distribution can be used to detect distinct groups of word–topic combinations.

However, dividing words into topics is not enough to fully capture the depth of the topics. Using certain words together does not encapsulate the context of the tweet. These words should be analyzed concurrently with the tweet–topic distribution to understand further the composition of each group.

LDA: Topic–Tweet Association

Within the framework of LDA, every tweet can be associated with topics based on γ (Figure 10). γ represents the probability that a certain tweet belongs in a particular topic. Figure 17 describes the Topic–Tweet distribution for the 1,131 relevant tweets over all 5 topics.

Figure 17: Topic–Tweet distribution for all 5 topics

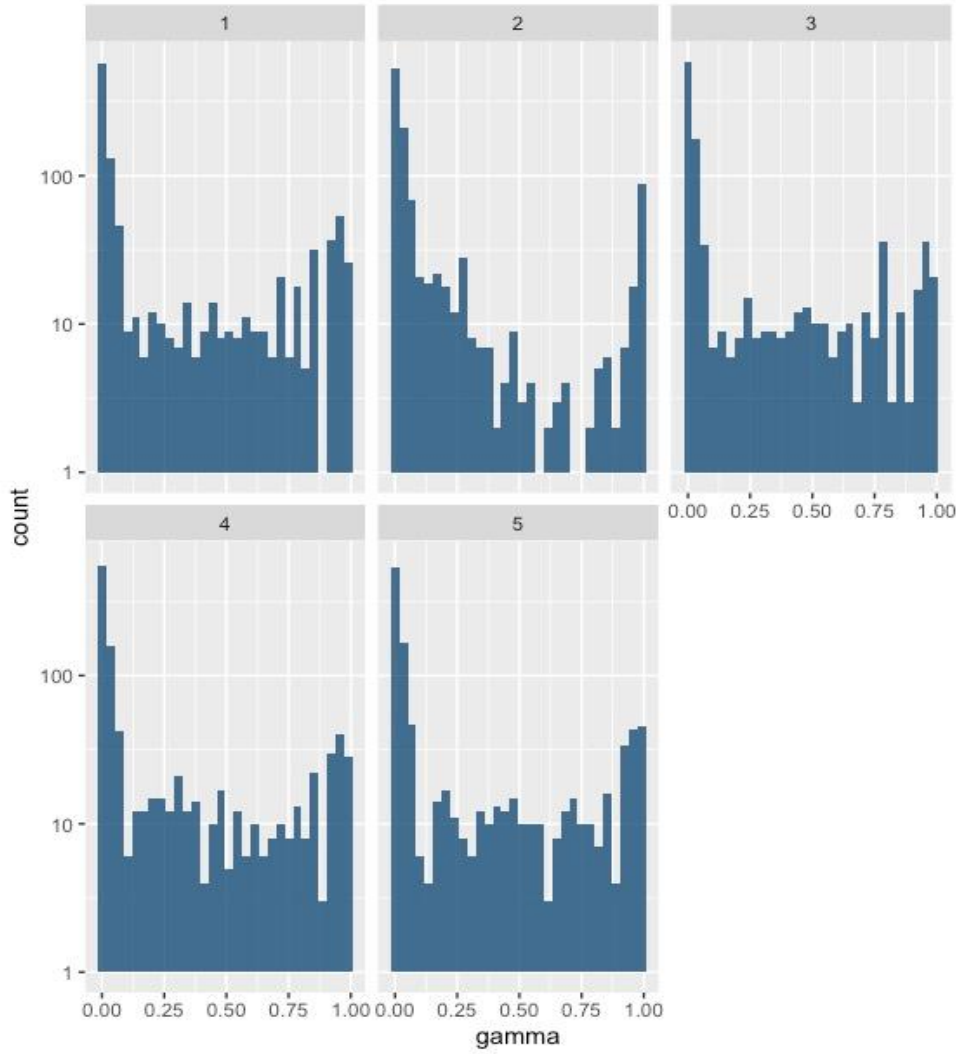


Figure 17 can be described as a histogram which represents the degree of intra-topic polarization. As γ probabilities represent topic–tweet associations, they are not mutually exclusive. Thus, each tweet can be described as being composed of multiple topics. To quantify the degree of polarization in each topic, a polarization index[43] is utilized. The measure of polarization, denoted by d is computed as follows:

$$d = \frac{|gc_u - gc_l|}{|\gamma_{max} - \gamma_{min}|}$$

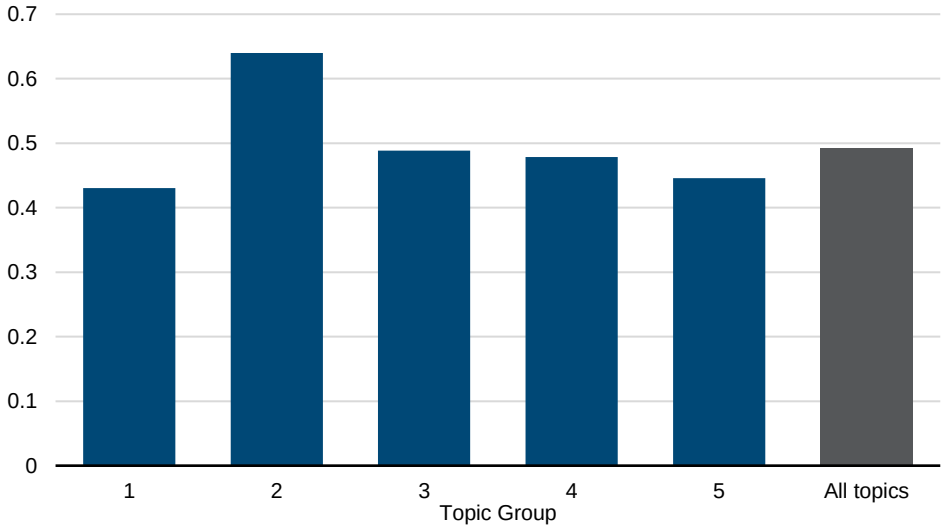
where:

$$gc_l = \frac{\int_0^{0.5} p(\gamma)\gamma \, d\gamma}{\int_0^{0.5} p(\gamma) \, d\gamma}$$

$$gc_u = \frac{\int_{0.5}^{1} p(\gamma)\gamma \, d\gamma}{\int_{0.5}^{1} p(\gamma) \, d\gamma}$$

The polarization index is a measure of normalized distance between the upper and lower gravity centres (gc_u and gc_l) of each topic. At extremes, the index yields d=0 where there is no seperation between gravity centres (the topic is described by identical tweets) and d=1 where there is maximal seperation between gravity centres (the topic consist of tweets that are at extreme ends and are completely and perfectly opposed).

Figure 18: Topic Polarization Index



As observed in Figure 18, topic 2 records the highest polarization compared to other topics. The polarization index for topic 2 is also significantly different from all other topics, which indicates the potential use of the LDA algorithm to classify tweets according to γ-values of topic 2. We will discuss further implications of this observation further in the next section.

[43] As proposed in Morales et al. (2015)

3.4. Similarities and Differences between Human Assessments and NLP

One other major use for tweet–topic association is to establish some relationship between our manual human assessments and natural language processing. Since γ-values are attached for each tweet and each tweet has its own categorical Information Disorder criteria based on human assessments, this section examines the corroboration between the two methods.

Tweet–Topic association also allows each tweet to be described as a composition of topics based on its semantic structure. Figure 19 through Figure 24 below summarizes the intersection between mean Tweet–Topic compositions of LDA and the results of human assessments.

Figure 19: Tweet–Topic composition (No Information Disorder)

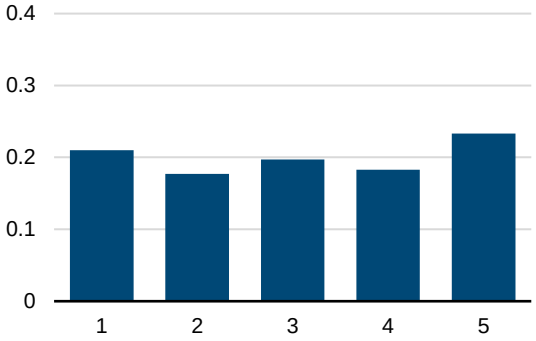


Figure 20: Tweet–Topic composition (Falseness)

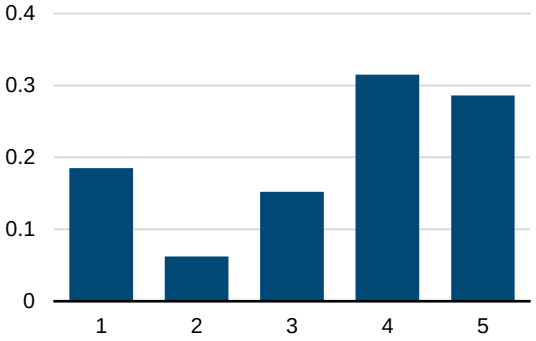


Figure 21:Tweet–Topic composition (Intent to Harm)

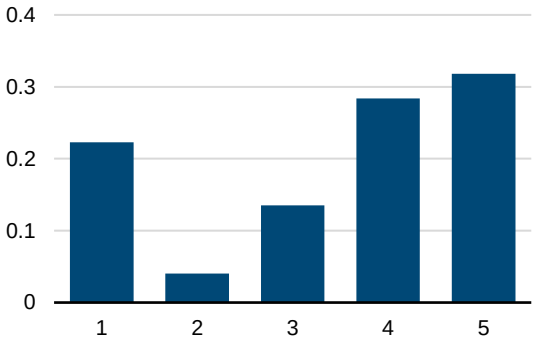


Figure 22: Tweet–Topic composition (Expressing Hate)

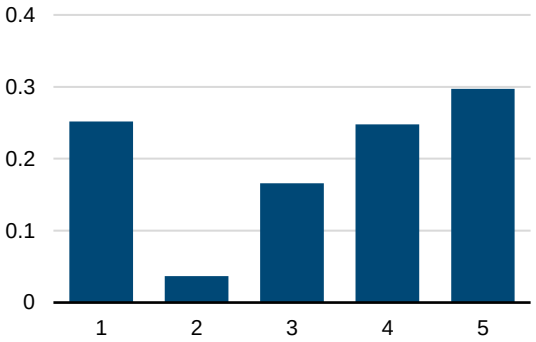


Figure 23: Tweet–Topic composition (Inciting Violence)

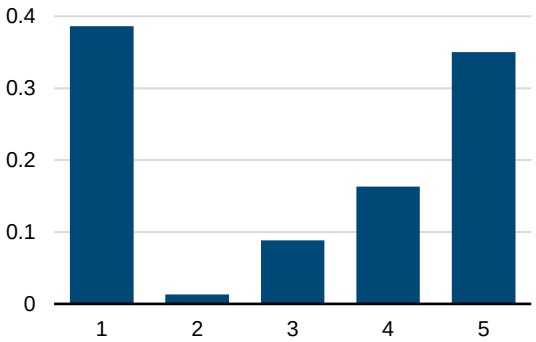
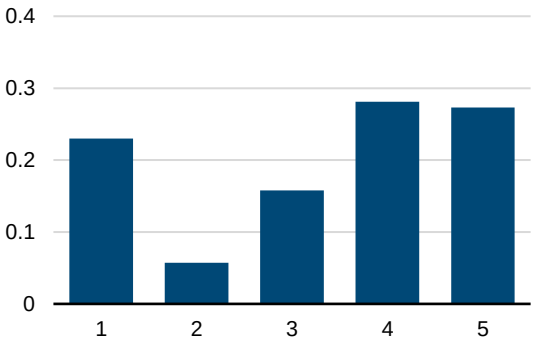


Figure 24: Tweet–Topic composition (Spreading Rumors)



It is observed from Figure 19 that the tweets which the coders identified to have no Information Disorder exhibits a more "uniform" distribution. Tweets that were identified as containing "Information Disorder" of the various categories did not exibit a uniform distribution as illustrated in Figure 20 to Figure 24.

Across all Information Disorder tweets, topic 2 has a mean γ-value close to zero. This means that a tweet with low to almost zero γ-value of topic 2 is a necessary factor for Information Disorder. The mean γ distribution for "spreading rumors", "intent to harm", and "expressing hate" is almost identical except with higher values for topics 4 and 5 for "intent to harm". The reason for the similarity could be attributed to the fact that the tweets that fall under these categories tend to be labelled together. "falseness" has a slight difference (topic 5 being slightly higher than the rest); this is because people tweeting false information might have done so without ill intentions. "Inciting violence" distribution shows a great disparity between topics 1 and 5 and with topic 2. The magnitude of mean γ-value for topic 5 is slightly above 0.4, topic 1 slightly above 0.3, and topic 2 almost zero.

Compared with "no Information Disorder" tweets, it seemed that the harmful category of tweets is linked with low mean γ-values for topic 2, and high γ-values for topics 1, and 5. High mean γ-values for topic 4 seemed to be linked closely with falseness and misinformation as it registered lower γ-values for "inciting violence".

However, topic 3's role in classifying Information Disorder remains ambiguous as it is almost the same across all categories. The vast difference between the mean γ-values of different topics shows that we can establish some distinction between the tweets using topic modeling via LDA.

While the results above do not show a one-to-one relationship between human assessments and natural language processing, the results suggest that there is a way to cluster Information Disorder using the various techniques documented here.

From our analysis, there is a clear difference in the semantic structure in sentences based on the bag-of-words model between different types of Information Disorder categories we have established in our earlier discussion. Tweet–topic distribution showed the algorithm's ability via LDA to detect semantic resemblance between different types of Information Disorder tweets. By describing the semantic structure of Information Disorder, the description may be further studied to better inform human assessments of Information Disorder.

3.5. Further improvements with more data and research

While our computational methods show promising results to detect Information Disorder in tweets, the next step to increase the efficiency for Information Disorder detection is through the use of sentiment analysis. Sentiment analysis analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language[44].

[44] Liu, n.d.

The core of sentiment analysis is the sentiment lexicon (a library of words with an attached sentiment rating); there are a few accessible lexicons available for non-commercial use such as the NRC Word-Emotion Lexicon Association[45] , the Bing lexicon[46], and the AFINN lexicon[47]; all however, are in English. The development of an easily accessible sentiment library in *Bahasa Melayu* is therefore an important element in Malay language opinion mining that could enable further opinion-based analysis to be conducted. Further research in sentiment analysis could prove to be useful for our study.

Our LDA analysis results are encouraging, but we can improve the quality of our results with more data. As a result of limitations of our access to Twitter APIs, we are only able to access just over 2,000 tweets with nearly half of them non-relevant to the subject of our study. A larger data set can contribute to a greater number of topics from LDA, giving us greater clarity in the distinctions between the topics.

While language is essential for information dispersion, there are also other important internet media besides text for the spread of information: most notably images, videos, and audio. Greater research efforts in natural language processing, image processing, speech recognition, and computer vision are needed to approach misinformation from different prongs of data analytics. The reason we could perform great analysis on language is because of its highly structured nature. Developing detection models of unstructured data such as images and videos of fake news are future areas to explore.

[45] Mohammad and Turney (2012)
[46] Liu (2010)
[47] Nielsen (2011)

4. Policy Considerations

The way we live and function in society is intrinsically tied to the use of language. The acceleration and democratization of communication in a highly digitalized world has revealed several vulnerabilities in the social fabric of countries and communities. While there are many policy considerations, we focus only on four policy recommendations which we believe could have the largest effects in combating Information Disorder.

1) Vulnerabilities emerging from a digitalized society can benefit from Augmented Intelligence

Firstly, without a means to first classify Information Disorder, it would be extremely difficult to detect and measure its spread on social media. For example, the Mueller Report[48] highlights, evidence of Russian involvement in US elections using social media as a tool to spread mis/dis/malinformation. This reveals the vulnerability of social media to a functioning democracy. Without a means to classify Information Disorder, there is no means to even detect activities such as disinformation campaigns. Our democracy is left vulnerable to manipulation both inside or outside the country.

However, with millions of tweets, Facebook posts, and other user generated media produced everyday, the prospects of having humans to go through and label them manually would be a futile task. Moreover, our case study indicates that while humans are able to consistently identify the presence of Information Disorder with respect to factual claims, the identification of "intent" proves to be more inconsistent.

The analysis from topic modeling indicates that LDA distinguishes factual claims and non-factual claims in similar ways when compared to human assessments. This has two potential implications:

1. LDA may be employed to statistically generate "features" of Information Disorder. These statistical features can then be employed as a filter to detect the presence of Information Disorder on a real time basis.
2. The use of LDA as a decision filter to first pre-select questionable media on social media platforms can potentially simplify the workload of fact-checkers.

Vulnerabilities emerging from a digitalized society require digitalized tools that can augment human decision making. Big data analytics have the potential to access different segments of society tweeting about the same news trend. Detailed data collection from social media—voluntary surveys of user information such as residence, income level, and education level—can allow future research to identify some key correlating factors that contribute to vulnerabilities towards misinformation and fake news. Big data analytics will allow us to study social trends with real time detailed data sets—which also comes with the need to invest in computer hardware

[48] Mueller (2019)

infrastructure as well as the human capital that could design and manage sophisticated algorithmic models in a relatively short amount of time.

2) Principles of objective fact-checking

Secondly, the role that reporters and journalists play in being the "fourth pillar of democracy" has been eroded by the advent of social media. Most Malaysians now heavily rely on Facebook and other social media platforms as their main news source[49]. Reporters and journalists hold to a set of standards when discharging their duties. However, with such a responsibility, they can also be sued and be held accountable for what they publish. Nonetheless, the advent of social media platforms have democratized the powers previously held by reporters and journalists, without the corresponding responsibilities of holding to standards of publishing. Reporters and journalists are no longer the monopoly in gatekeeping the diffusion of information in this day and age.

As pointed out by [50], there are many weaknesses with regards to *Sebenarnya.my's* effectiveness. For example, *Sebenarnya.my* relies a lot on "tips" and has a heavy emphasis on government agencies as a source of truth. It does not have a robust fact-checking environment in which a diverse membership can contribute towards better fact-checking a certain media.

While government efforts and intentions to combat Information Disorder have been observed in this pandemic, the means through which "questionable media" is classified is not. [51] also highlights that the proximity of *Sebenarnya.my* to authorities could undermine the perceived validity of their exercise, should trust in government erodes.

The International Fact-Checking Network (IFCN)[52] outlines the following principles that are foundational to fact-checking:

1. A commitment to **Non-Partisanship and Fairness**
    - Claims are fact-checked using the same standards for every fact-check. All sides are taken into account. Evidence dictates conclusions.
2. A commitment to **Standards and Transparency of Sources**
    - All sources are published so that readers are able to verify and externally validate the findings.
3. A commitment to **Transparency of Funding and Organization**
    - Fact-checkers reveal their source of funding, qualifications and affiliations.
4. A commitment to **Standards and Transparency of Methodology**
    - Fact-checkers explain the methodology that they use to select, research, write, edit, publish and correct their fact checks.
5. A commitment to an **Open and Honest Corrections Policy**

[49] Vase.ai (2019)
[50] Harris and Farlina (2020)
[51] Harris and Farlina (2020)
[52] "IFCN Code of Principles" (n.d.)

- Fact-checkers publish their corrections policy, correct clearly and transparently in line with the policy.

In this regard, the percieved trustworthiness of *Sebenarnya.my* could be improved by making some or all of these dimensions explicit. The goal of making this information accessible, is to indicate to readers that fact-checkers are credible individuals, that perform fact-checking in a non-biased way, using a consistent method to arrive at conclusions.

3) Civic responsibilities of handling Information Disorder

Thirdly, lies spread much faster, deeper and wider as compared to truth. In a mass study[53] which looked at 126,000 verified stories tweeted by over 3 million people, from the inception of twitter in 2006 to 2017:

1. Truth rarely diffused to more than 1,000 people, while the top 1% of false-news cascaded to between 1,000 and 100,000 people.
2. Truth took about 6x as long as falsehood to reach 1,500 people, and 20x as long to reach a cascade depth[54] of 10.
3. Truth never exceeded a cascade depth of 10, while falsehood reached a depth of 19 nearly 10 times faster than truth reaching a depth of 10.

While not explicitly studied in our paper[55], the characteristics of falsehood diffusion brings into question the role that *Sebenarnya.my* and other government initiatives play in combating Information Disorder in Malaysia. In particular, even if *Sebenarnya.my* can hypothetically detect and classify all "compromised" media, the dissemination of its results might not be as deep, as wide or as fast as the spreading of "false news", whose damage might have already been done.

In an ideal world, to effectively combat the spread of "false news", the results of fact-checking have to be seen by all individuals who have potentially viewed the "false news". To inch closer to this ideal, *Sebenarnya.my* has to be empowered to work together with social media platforms to first flag "false news"; and to disseminate the results of the fact-checking exercise in a timely manner to all platform users who might have viewed or interacted with questionable media.

Recently in the 2020 US elections, Twitter has taken action to censor tweets with false information regarding the election[56]; they went as far as to remove President Trump's Twitter account. Twitter's role in regulating social media is the result of multiple collaborations with various news agencies from all over the political spectrums: from Fox News to CNN to Associated Press (a non-partisan not-for-profit news agency). Therefore, regulating social media content requires a decentralized responsible collaboration from a diverse set of bodies (government, NGO, private sectors, news agencies).

---

[53] Vosoughi, Roy, and Aral (2018)

[54] Cascade depths refer to independent sharing of unbroken retweet chains with a common, singular origin.

[55] We intend to study this phenomena in a future publication.

[56] Gadde and Kayvon Beykpour, n.d.

4) Digital and Online Media Literacy

Regulating information and fact checking are highly reliant on the choices that people make whether to believe or disbelieve the information that is presented to them. The spread of misinformation at the end of the day is the result of willing actors propagating the message[57]. Therefore, combating misinformation should also factor in user participation—which can be approached from two paths: critical thinking and institutional trust.

Digital literacy and the ability to discern false information on the Internet requires critical thinking and source evaluation. Therefore, civic education in schools and in public should emphasize the role of individuals in society and their responsibilities to be informed citizens. School curriculums should prioritize the role of critical thinking in education. In 2015, Stanford History Education ran a study on under-resourced schools in Los Angeles and Minneapolis suburbs and found that the student's ability to reason from online information is "bleak"[58]. Introducing online literacy educational programmes in schools and to the wider general public especially underserved communities is a way forward to foster an online-intelligent society.

Reducing the trust gap among people and government institutions is crucial to minimize the effect of Information Disorder. Low institutional trust was shown to have an effect in the lower likelihood to adopt preventative behaviors during the Ebola outbreak[59]. The study showed that greater institutional mistrust is correlated with widespread misinformation (belief that Ebola was not real was prevalent) causing behaviors such as refusal to vaccinate or seek medical assistance and lower compliance to messages from authorities, increasing the risk of spread of the Ebola virus. Therefore, public confidence in institutions is essential to stem the prevalence of Information Disorder and minimizing its harmful real-world effects.

A 2020 study by Ipsos found that 59% of Malaysians do not trust the government[60]. Although decreasing that trust gap is not within the scope of the study, one of the ways to improve government trust is to improve e-government facilities[61]. Institutional-based trust; commitment to transparency and its responsibility for its citizens; and process-based and institutional-based trust; government efficiency and grassroots participation; are areas that need to be reevaluated. There is an urgent need for stronger collaboration and relationship between the government and its people. Closing the trust gap contributes to the effectiveness for authorities to spread important information without becoming fake news.

[57] Marwick (2018)
[58] Wineberg et. al (2016)
[59] Vinck et.al (2019)
[60] Ipsos (2020)
[61] Tolbert and Mossberger (2006)

5. Conclusion

"That which is measured improves. That which is measured and reported improves exponentially." – Karl Pearson

Measurement is at the heart of every data-generating process. It is the starting point that enables discourses to flourish. There is however a sinister side to Karl Pearson's quote—"That which is measured can often be forged or fabricated". The vulnerabilities that emerge with the use of social media is ultimately a function of un-truth and of malicious intent.

To sustain an enabling policy environment to combat misinformation and safeguard society against these vulnerabilities, governments, corporations and society must commit to uphold truth as our "sacred value", as opposed to "social justice" or "political correctness" for example, even though these are quite often not mutually exclusive.

The fourth pillar of democracy requires the commitment to uphold truth above all else, in order for other features (like "social justice") to play a positive role in society.

6. Appendices

6.1. Appendix A: Definitions of Bahasa Melayu Words

Appendix A provides a simple one-to-one translation of Bahasa Melayu words in this discussion paper for non-Malaysian readers.

Word	Translation	Word	Translation	Word	Translation
aduhlah*	Oh no!	keseluruhan	overall	positif	positive
adumak*	Oh no!	kesemuanya	in total	punca	source
akan	will, cause	kesihatan	health	ramai	many
atau	or	ketua	leader	rapat	close
baharu	new	kluster	cluster	rasmi	official
balik	return	kontak	contact	rentetan	string, chain
baru	new	lapangan	field	rujuk	refer
berjalan	walking	laporan	report	sebanyak	as many as
berumur	aged	lelaki	man	sebulan	a month
bulan	month	lima	five	sedang	currently
dapat	got	luar	outside	sejarah	history
dicaj	charged	mempunyai	have	sembuh	recover
dijalankan	executed	mendadak	suddenly	semua	all
dilaporkan	reported	mengesahkan	validate	senarai	list
dimaklumkan	informed	naik	up	siapa	who
dirawat	treated	nak*	want, child	siasat	investigate
disahkan	validated	negara	country	situasi	situation
empat	four	oleh	by	stabil	stable
gelombang	wave	orang	person	sumber	source
halahaii*	(sigh)	org*	person	tak	not
hingga	until, till, to	pasukan	team	takpelah*	it's ok
jadi	become	pengarah	director	tempatan	local
kedua	second	pengesanan	detection	terkini	latest
kekal	remain	pergi	go	termasuk	including
kementerian	ministry	perjalanan	trip	thn*	year
kena	affected, hit	perkembangan	progress	tujuh	seven
keputusan	results	pertengahan	middle	tunggu	wait
kes	case	pesakit	patient	wanita	woman

Note: * Refers to shorthand and abbreviated words

6.2. Appendix B: Kes-26 Public Letter

STATEMENT BY HISHAM HAMDAN, COVID-19 PATIENT

First and foremost, I would like to record my deepest appreciation for the incredibly diligent and dedicated folks across the healthcare system in Malaysia who are doing a tremendous job in this very challenging period. Their work ethic and dedication to the cause are second to none.

From the medical professionals at the Subang Jaya Medical Centre ("SJMC"), to the doctors at Sungai Buloh Hospital, led by Dr. Yasmin Mohd. Ghani, who have been working extremely hard to treat my condition –as well as the conditions of all the other COVID-19 patients –to the officers at the Ministry of Health, it has made me extremely proud to know that Malaysia has a wonderful healthcare system. In particular, I would like to especially commend Dr. Muhammad Haikal Ghazali from Selangor Health State Department and Dr. Zaza Rida Zakiman from the Petaling Health District Office.

Next, I would like to take this opportunity to address some of the media reports that have been circulating with regards to my case, Case 26. I believe that it is important for me to share the facts regarding my particular circumstance so that the public has a clear picture of events.

On the 27th of February 2020, I started exhibiting symptoms, namely fever and a cough. That afternoon, I went to the SJMC Outpatient Center to get myself tested as I was concerned that I had dengue fever. While there, I also specifically requested for the COVID-19 test. After doing the test, I went home and stayed home. On the 28th of February 2020, in the evening, I received my first round of test results stating that I had tested positive. I was then asked to proceed to Sungai Buloh Hospital on the 29th of February to be isolated and treated. It was there that my positive results were confirmed.

At that point, I was the 26th person in Malaysia to be tested positive for the COVID-19 disease, which does not necessarily mean that I was the 26th person in Malaysia to be infected by it. There were potentially others whohad been infected earlier but not tested. Accordingly, I worked with Dr. Haikal and Dr. Zaza to come up with a Contact Tracing list, along with colleagues at UDA and at Khazanah. In addition, my family was also tested.

My family have all, Alhamdullilah, tested negative. It is hugely unfortunate that two individuals have caught the COVID-19 from me, namely my driver at UDA as well as the SJMC paramedic who was treating me. They are, at present, being treated with the utmost care and professionalism from Malaysia's healthcare professionals.

The second wave of COVID-19 cases in Malaysia are linked to me, that I think is clear. But being linked to me and having originated from me are two entirely different things. The Ministry of Health is still working diligently and must be applauded for continuing to search for Patient Zero. I was at several meetings from the 21st to the 27th with individuals who have since been confirmed positive for the COVID-19 disease. As I mentioned, the Ministry of Health is still searching for Patient Zero. I just happened to be the first person who was tested from this string of meetings. At the same time, it is worth noting that there were certain meetings on the morning of the 24th of February where all 13 non-UDA board and management individuals all tested negative. I also did attend a Ministry function in my capacity as Chairman of UDA on the 27th of

February, but I would like to clarify that I did not attend any political functions. At this stage, I must also commend Dr. Haikal andDr. Zaza for arranging a sizeable number of tests for those in UDA and Khazanah who were in close contact with me.

The next point I would like to touch on is my visit to Shanghai. I was in Shanghai attending a conference from the 13th to the 17th of January. There are two issues to raise here. Up to that point, the only confirmed cases in China were from the city of Wuhan, which I never visited. There was no suspicion, at that time, for any concern with regards to visiting Shanghai. Shanghai's first recorded case was on the 20th of January. In Malaysia, on the 25th of January, a week after I arrived home from Shanghai, the Ministry of Health issued an advisory for Malaysia to postpone or avoid travel to China. On the 30th of January, the World Health Organisation, as a result of the novel coronavirus, officially declared a "public health emergency of international concern." This all happened well after I returned from Shanghai.

The second issue is that all scientific and medical research we know so far points to the virus having a two-weekincubation period. Given that I returned on the 17th of January, and given that I exhibited symptoms on the 27th of February, it is –as far as medical research is concerned –not possible for me to have obtained the virus from my trip to Shanghai.

Furthermore, the earliest close contact patients linked to me are from a meeting on the 21st of February, 5 weeks after I returned. Unless new medical research tells us otherwise, it is important to keep the facts, as we know them now, clear –my visit to Shanghai is not linked to my positive confirmation.

To summarise, based on the facts that I have laid out, while it is true that I am linked to the second wave of cases, being linked to and being the source of are two entirely different things. We need to give our full support to MOH as they identify Patient Zero. Next, given what medical research tells us, the incubation period for the virusis two weeks, and so, I did not catch it from my trip to Shanghai. Furthermore, at the time of my visit to Shanghai, there were no recorded cases there as yet.

Finally, I trust that we will all be guided by the facts as well as the exemplary work done bythe entire medical professionals at the Ministry of Health. I thank them all for their service to our country. I also call on the public to give them their full support, to use facts before spreading news or opinions, and to respect the confidentiality and privacy of individuals going through their recovery. I would like to also send out my prayers to all UDA staff and their families, as well as all other infected patients, whether in Malaysia or globally, that are still under treatment as I understand howdifficult it is and wish them the speediest of recoveries.

Thank you.

HISHAM HAMDAN
SUNGAI BULOH
6 MARCH 2020

7. References

Bakshy, E., S. Messing, and L. A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348 (6239):1130–32. https://doi.org/10.1126/science.aaa1160.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan):993–1022.

Caroline Tolbert and Karen Mossberger. 2006. "The Effects of E-Govenment on Trust and Confidence in Government." *Public Adminstration Review*, May, 354–69.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems* 22:288–96.

DOSM. 2020. "Current Population Estimates, Malaysia, 2020." July 15, 2020. https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=155&bul_id=OVByWjg5YkQ3MWFZRTN5bDJiaEVhZz09&menu_id=L0pheU43NWJwRWVSZklWdzQ4TlhUUT09.

Gadde, Vijaya and Kayvon Beykpour. n.d. "Additional Steps We're Taking Ahead of the 2020 US Election." Twitter.

Guterres, A. 2019. "United Nations Strategy and Plan of Action on Hate Speech." *Taken from: Https://Www. Un. Org/En/Genocideprevention/Documents/U*, no. 20Strategy.

Harris, Zainul, and Said Farlina. 2020. *The COVID-19 Infodemic in Malaysia*. Policy Paper. Institute of Strategic and International Studies Malaysia. https://www.isis.org.my/2020/08/24/the-covid-19-infodemic-in-malaysia-scale-scope-and-policy-responses/.

"IFCN Code of Principles." n.d. Accessed November 16, 2020. https://www.ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles.

Ipsos. 2020. "Do Malaysians Lack Trust in Government and Institutions?" Press Release. Kuala Lumpur, Malaysia: Ipsos Sdn Bhd.

Ireton, Cherilyn, and Julie Posetti. 2018. *Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training*. UNESCO Publishing.

"KKM Portal MyHealth on Twitter." n.d. Twitter. Accessed February 22, 2021. https://twitter.com/MyHEALTHKKM/status/1235880123384991744.

Liu, Bing. 2010. "Sentiment Analysis and Subjectivity." In *Handbook of Natural Language Processing*, Second, 629–61. Machine Learning and Pattern Recognition Series. Cambridge, UK: CRC Press.

———. n.d. *Sentiment Analysis and Opinion Mining*. Syntheis Lectures on Human Language Technologies 16. Morgan and Claypool Publishers.

Malaysian Communications and Multimedia Commission. 2019. "Internet Users Survey 2018." Annual Survey. Malaysia. https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/Internet-Users-Survey-2018.pdf.

"Managing the COVID-19 Infodemic: Promoting Healthy Behaviours and Mitigating the Harm from Misinformation and Disinformation." n.d. Accessed October 20, 2020. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-

promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation.

Marwick, Alice E. 2018. "Why Do People Share Fake News? A Sociotechnical Model of Media Effects." *Georgetown Law Technology Review* 2 (2):474–512.

McLuhan, Marshall, and Quentin Fiore. 1967. "The Medium Is the Message." *New York* 123:126–28.

McLuhan, Marshall, and MARSHALL AUTOR MCLUHAN. 1994. *Understanding Media: The Extensions of Man*. MIT press.

Mitra, Tanushree, and Eric Gilbert. 2015. "Credbank: A Large-Scale Social Media Corpus with Associated Credibility Annotations." In *ICWSM*, 258–67.

Mohammad, Saif, and Peter Turney. 2012. "Crowdsourcing a Word-Emotion Association Lexicon." *Computational Intelligence*, September.

Morales, Alfredo Jose, Javier Borondo, Juan Carlos Losada, and Rosa M. Benito. 2015. "Measuring Political Polarization: Twitter Shows the Two Sides of Venezuela." *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25 (3). AIP Publishing LLC:033114.

Mueller, Robert S. 2019. *The Mueller Report: Report on the Investigation into Russian Interference in the 2016 Presidential Election*. WSBLD.

Ngah, Nazura. 2018. "FAQs: What You Need to Know about the Anti-Fake News Bill 2018." NST Online. March 26, 2018. https://www.nst.com.my/news/nation/2018/03/349691/faqs-what-you-need-know-about-anti-fake-news-bill-2018.

Nielsen, Finn. 2011. "A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs." *ArXiv*, March.

Rio, Victoire. 2020. "The Role of Social Media in Fomenting Violence: Myanmar." *Toda Peace Institute* Policy Brief No. 78 (June).

Sarawak Report. 2020. "EXCLUSIVE - Health Rules Demand PM8 And All Co-Conspirators Must Immediately Self-Isolate For 14 Days?" Sarawak Report. March 1, 2020. http://www.sarawakreport.org/2020/03/health-rules-demand-pm8-and-all-co-conspirators-must-immediately-self-isolate-14-days/.

Sarkar, Dipanjan. 2016. "Text Analytics with Python." Springer.

Sen, Amartya Kumar. 1999. "Democracy as a Universal Value." *Journal of Democracy* 10 (3). Johns Hopkins University Press:3–17.

Vase.ai. 2019. "Malaysia's 2019 Media Consumption Report." Learning Resources | Vase Actionable Intelligence. September 6, 2019. https://vase.ai/resources/malaysias-media-consumption-2019/.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The Spread of True and False News Online." *Science* 359 (6380). American Association for the Advancement of Science:1146–51. https://doi.org/10.1126/science.aap9559.

Wardle, Claire, and Hossein Derakhshan. 2017. "Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking." Council of Europe report DGI. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Wittgenstein, Ludwig. 1921. "Logisch-Philosophische Abhandlung." *Annalen Der Naturphilosophie* 14:185–262.

———. 2009. *Philosophical Investigations*. John Wiley & Sons.

Wittgenstein, Ludwig, and Luiz Henrique Lopes dos Santos. 1994. *Tractatus Logico-Philosophicus*. Edusp.

Wu, Liang, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. "Misinformation in Social Media: Definition, Manipulation, and Detection." *ACM SIGKDD Explorations Newsletter* 21 (2). ACM New York, NY, USA:80–90. https://doi.org/10.1145/3373464.3373475.