

Unmasking Fake News: Lexical Patterns in COVID-19 Misinformation

Sharmin Haque and Adlina Ariffin¹

International Islamic University Malaysia

ABSTRACT

The surge of fake news on social media during the COVID-19 pandemic posed serious risks to public health and social stability. While previous studies have largely focused on computational detection of misinformation, less attention has been paid to identifying human-recognisable linguistic features that enable real-time fake news detection. This exploratory study investigated the lexical features that characterise fake COVID-19 news and identified the most distinctive features that differentiate fake news from authentic reporting. Using a qualitative research design, this study analysed 27 misinformation texts to examine a corpus of English-language fake COVID-19 news articles collected from an International Fact-Checking Network (IFCN)-certified fact-checking website: SEBENARNYA.MY (Malaysia). The findings may suggest preliminary evidence of linguistic patterns of recurring lexical features in fake news, including random capitalisation (capitalised words and all-uppercase text), irregular use of lowercase letters, repetition of content words and phrases, multiple types of spelling errors, informal shortened words, inappropriate lexical choices, and non-standard formatting such as excessive bold text and improper word spacing. The results indicate that fake news shows orthographic instability, reduced lexical diversity, and register mismatch, reflecting deceptive intent and cognitive constraints associated with fabricated narratives. This study contributes to fake news research by foregrounding interpretable, surface-level lexical cues that enable readers to identify misinformation without reliance on automated tools. The findings have important implications for media literacy, public health communication, and linguistic approaches to detect misinformation.

Keywords: *Corpus analysis, COVID-19, Fake news, Lexical features, Social media.*

INTRODUCTION

News plays a vital role in helping us understand society and the world around us. Traditionally, news has been regarded as a reliable source of information, primarily produced by professional journalists. However, this role has been increasingly challenged by the rise of digital media (Anderson, 2025). Over time, news has evolved from handwritten newsletters and pamphlets in the 18th century, to printed newspapers in the 20th, and eventually to digital platforms with real-time publications on websites, blogs, and social media in the 21st century.

Audiences have also shifted from passive consumers to active creators. In short, anyone can publish anything online, including false stories or fake news. This blurs the line between journalism and opinionated information (e.g., fake news), profoundly transforming the

¹ The corresponding author can be contacted at the Department of English Language and Literature, AbdulHamid AbuSulayman Kulliyah of Islamic Revealed Knowledge and Human Sciences, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia (email: adlina@iium.edu.my)

fundamental news ecosystem (Medjanova & Atayeva, 2024). Nowadays, fake news is spreading rapidly on digital media and misleading us from the truth.

A study by Surjatmodjo et al. (2024) shows that on social media, fake news spreads roughly six times faster than credible information. Thus, there is a great need for its detection. The reach and large circulation of fake news can be better understood through some recent examples, such as coronavirus and COVID-19 vaccine-related infodemics and conspiracy theories across the world (Rodrigues & Xu, 2020; Islam et al., 2021); the spread of fake news amid several presidential elections worldwide, for example, the “Pizzagate” scandal and “Hillary Health Scare” stories during the 2016 US presidential elections (Stabile, Grant, Purohit, & Harris, 2019); the 2018 Malaysia 14th general election and the 2019 Indonesia presidential election (Jalli & Idris, 2019); the 2017 South Korean presidential election (Go & Lee, 2020); fake stories about the Hong Kong protests in June 2019 (Zervopoulos, Alavanou, Bezas, Papamichail, Maragoudakis, & Kermanidis, 2020); the spread of lies using fake and automated Facebook and Twitter accounts and bots by an Indonesian-based cyber network, “Muslim Cyber Army” (Lamb, 2018); and fake news about climate change (Treen, Williams, & O’Neill, 2020), to name a few.

COVID-19-related fake news includes fake cures, such as the claim that injecting bleach is a cure (WHO, 2020), false theories such as the claim that the coronavirus was engineered in a Wuhan lab (Andersen, Rambaut, Lipkin, Holmes, & Garry, 2020; Cohen, 2020), and the rumour that the 5G mobile network is responsible for the coronavirus disease (BBC News, 2020). It is also worth noting the film ‘Plandemic’, which contains false claims about the coronavirus and public health (e.g., that wearing a mask may activate the coronavirus) and received millions of views and spread widely shortly after it was aired online in May 2020 (Cook, van der Linden, Lewandowsky, & Ecker, 2020).

Anti-vaccine groups are now spreading COVID-19 vaccine-related fake news and misinformation on digital media (Kanozia & Arya, 2021). In real life, COVID-19 vaccine hesitancy is now being seen across countries (Bono et al., 2021; Kanozia & Arya, 2021). Similar patterns emerged during the 2016 U.S. presidential elections. Nineteen major news websites posted 7,367,000 fake stories, which were shared, reacted to, and commented on via social media (Silverman, 2016). This had a significant impact on politics, creating political polarisation that influenced voters’ choices of candidate (Amin, 2018).

This gives rise to a question: what is the novelty of fake news that makes people believe them? The novelty of fake news lies in its language. Because fake news is generated to deceive people, and because the language (linguistic features) of a text varies by purpose (Alvestad et al., 2020), fake news should contain certain distinguishable linguistic features. For example, it may use words associated with basic human emotions such as surprise, disgust (Vosoughi et al., 2018), and anger, conveying urgency and commitment. These elements are meant to capture the audience’s attention quickly and propel them to believe it. Put differently, fake news consists of tricky, deceptive words and language. It is designed to persuade readers; in other words, it is a linguistic trap (Luu, 2019). These explanations require an investigation into the linguistic features of fake news detection. The concept of fake news is being studied, particularly, from two strands, namely: (1) how fake news spreads and (2) the contents of fake news (Horne & Adali, 2017).

This study focuses on the content of fake news. Unlike earlier work, this study examines fake news through the lens of linguistics and socio-psychology. It uses textual feature analysis and a comprehensive socio-psychology survey. This helps the researcher better understand the

linguistic patterns and anomalies of fake news, facilitating the development of a linguistic framework that helps ordinary people improve their ability to check news veracity manually without complex fact-checking tools.

Currently, research in fake news detection mainly focuses on using technology as the core component. For this purpose, two main types of tools are used: detection systems that can automatically identify misinformation, and fact-checking technologies used to verify content authenticity. Undoubtedly, these tools play a significant role when put into use; however, to function properly, they rely on three supporting elements: first, well-trained algorithms; second, various pre-written and pre-set identification features; and finally, third-party institutions unrelated to the parties involved to conduct the final verification. They cannot operate without any one of these elements. Precisely because of these numerous limitations, the shortcomings of these tools are fully exposed as soon as they encounter new situations. If a completely new linguistic expression pattern that has never been seen before emerges online, they are completely unable to identify it. What is worse, when misinformation first appears online, and people urgently need to judge its authenticity to make immediate decisions, these tools are completely useless. (Zhang & Ghorbani, 2020; Koch et al., 2023).

Comparatively little research has investigated interpretable lexical cues that ordinary readers can recognise without computational assistance. The current exploratory study attempts to fill these gaps by analysing the lexical features of 27 COVID-19 misinformation texts in English drawn from SEBENARNYA. MY, the official fact-checking portal of Malaysia, which is accredited by IFCN (International Fact-Checking Network). In contrast to existing work that mainly assesses computational models or large-scale international databases, this work leverages in-depth qualitative analysis of a Malaysian corpus to reveal human-recognisable lexical features related to misinformation. This study contributes to the area of linguistics for misinformation detection despite being a country-specific perspective from Malaysia, where readers can directly observe surface-level lexical cues. This paper uses a systematic textual analysis to respond to the following research questions:

- i. What lexical features characterise fake COVID-19 news?
- ii. Which lexical features are most prominent in distinguishing fake COVID-19 news from real news?

This research, foregrounding interpretable and reader-oriented lexical cues, aims to make public media literacy easier by enabling fake news identification without reference to artificial intelligence tools in real-time and putting technology firmly in the back seat. This approach should help develop safer and more resilient digital information environments.

METHOD

Research Design

The current qualitative research has followed a three-step process, and these steps include corpus compilation (step one), data set preparation (step two), and lexical feature analysis through textual analysis (step three). The following figure 1 depicts the research procedure.

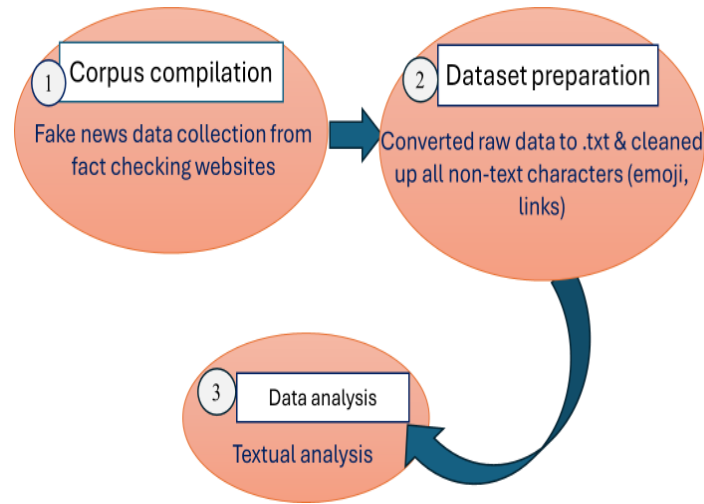


Figure 1. Overview of the research procedure

Conceptual Framework

The theoretical background of the research lies at the intersection of forensic linguistics, linguistic feature analysis, and the theory of Unnddeutsch, which explains the ways language can be applied in separating truth from lies. The above-mentioned theoretical approach helps in identifying lexical peculiarities of fake news during the COVID-19 pandemic.

Forensic Linguistic Analysis of Fake News

Language is integral to committing different types of crimes, including threatening, bribery, blackmail, extortion, and perjury, which all hinge on manipulation through the deliberate use of language (Sousa-Silva, 2022). Fake news is a linguistic crime as it involves the use of language with the aim of misleading and manipulating an audience (Tiersma & Solan, 2012; Sousa-Silva, 2022). The field of forensic linguistics regards language as evidence, where forensic methods are used to conduct linguistic analyses in a bid to identify patterns suggesting manipulation or deception. In relation to fake news detection, forensic linguistic analysis offers an effective way to analyse language by looking at aspects like vocabulary usage, linguistic structure and discourse to be able to separate fake news from genuine news. In view of this, using forensic linguistic concepts in detecting fake news about COVID-19 allows for the current study to look at language not just as a means of communication but also as linguistic evidence.

Linguistic Feature Analysis in Fake News Detection

Scholars who study actual communication situations have paid increased attention to the linguistic signals used to determine falsification of the message. It is generally acknowledged that linguistic elements differ in relation to the communicative situation, its purposes, and communicative functions (Biber & Conrad, 2019). Such variations help to evaluate the credibility of information. According to empirical studies related to fake news, such news possesses specific linguistic markers. For instance, Mahyoob et al. (2021) demonstrated that fake social media news contained more frequent use of reported speech, passives, negations, conjunctions, hedges, quotes, superlatives, and expressions of uncertainty, as well as less frequent use of proper names, comparative adjectives, and syntactically correct

sentences. Similarly, Grieve and Woodfield (2023) identified differences in density, confidence, and linguistic accuracy between fake and authentic news when analysing Jason Blair's stories. These studies confirm the primary assumption that fake news has linguistic peculiarities that careful analysis can detect. Thus, the importance of analysing lexical characteristics (word frequency, vocabulary richness) as reliable indicators of falsification is proven.

Integration of the Unndeutsch Hypothesis

One of the most important theoretical frameworks supporting this research is the Unndeutsch Hypothesis, developed in forensic psychology. According to this hypothesis, a narrative derived from real-life experience differs systematically in its properties from a fabricated one. That is, true and fabricated narratives draw on distinct linguistic constructions and mental processes. Although it applies to textual material, the Unndeutsch Hypothesis implies that news reports based on real-life events are more detailed and consistent, contextually embedded, and linguistically coherent. Meanwhile, fabricated news reports are expected to show contradictions, lower information density, more subjective and vague language, and linguistic techniques aimed at persuading rather than describing the situation. This hypothesis aligns with and theoretically strengthens the current study's findings from the perspective of forensic linguistics and corpus analysis. It provides a cognitive-linguistic explanation for how fake and real news differ lexically and stylistically.

Corpus Construction

The corpus was compiled from SEBENARNYA.MY, Malaysia's official fact-checking portal and a verified signatory of the International Fact-Checking Network (IFCN). The website verifies misinformation circulating through social media, messaging applications, and online news platforms. All selected misinformation texts were publicly accessible and were published between March 2020 and July 2021, covering the most critical period of the COVID-19 pandemic in Malaysia.

Sampling Criteria

The corpus was compiled using purposive sampling to ensure that all texts were relevant to the research objectives. The following inclusion criteria were applied:

- i. The articles were identified and verified as false by SEBENARNYA.MY, an IFCN-certified fact-checking organisation.
- ii. The articles were related to COVID-19.
- iii. The articles were written in English.
- iv. The articles were published between March 2020 and July 2021, corresponding to a major period of COVID-19 misinformation in Malaysia.
- v. Only one version of each misinformation item was included to avoid duplication.

The following exclusion criteria were applied:

- i. News items written in languages other than English.
- ii. Duplicate or substantially similar misinformation reports.
- iii. Articles containing only images or videos without sufficient textual content for linguistic analysis.
- iv. Updates or corrections referring to previously analysed misinformation items.

Corpus Size

The final corpus consisted of 27 English-language misinformation texts, comprising 8450 word tokens and 2310 word types. Reporting the corpus size provides an indication of the linguistic data analysed and facilitates comparison with future corpus-based studies.

Unit of Analysis

The unit of analysis in this study was the individual lexical feature occurring within each misinformation text. The lexical features entailed words, word forms, spelling features, orthographic rules, and text format that could be systematically analysed (Baxter, 2020). Every time a lexical feature appeared in the text, it would be coded based on a predetermined code book, without considering how many times it appeared.

Coding Framework for Lexical Features

Lexical Feature	Operational Definition
Capitalisation	Use of ALL CAPS or unnecessary capital letters for emphasis.
Lowercase irregularity	Inappropriate use of lowercase letters where standard English requires capitals.
Repetition	Repeated use of identical words or phrases to increase emphasis.
Omission spelling	Missing letters in words (e.g., "goverment").
Non-word spelling	Misspelled words that do not exist in English.
Homophonic spelling	Words replaced by similar-sounding words (e.g., "their" for "there").
Incorrect kerning	Unusual spacing within words or between letters.
Scriptio continua	Words written without appropriate spacing.
Informal shortened words	Abbreviated or non-standard lexical forms commonly used online.
Inappropriate lexical choice	Word selection inconsistent with standard news reporting.
Excessive bold formatting	Overuse of boldface to attract attention.
Improper spacing	Irregular spaces before, after, or within words.

Coding Procedure

The coding procedure entailed four steps. Firstly, all the texts were downloaded into Microsoft Word. Secondly, all the texts associated with the misinformation articles were reread to ensure that the researcher is acquainted with their linguistic properties. Thirdly, all lexical properties were identified based on the definitions established during the forensic linguistic and corpus linguistic literature reviews. Lastly, all the cases of a lexical phenomenon were coded manually, using a coding sheet. Similar linguistic phenomena were categorised under larger categories, while ambiguous cases were reassessed in light of the definitions of coding. Manual coding was preferred because certain lexical properties could not be automatically identified by the corpus software.

Reliability of Coding

The coding framework was examined before the analysis to maintain the validity of the qualitative analysis process. After the first coding stage, all documents were reviewed again in a second coding stage about two weeks later. The team resolved differences between the first and second coding stages using the operational definition of the variables.

Ethical Considerations

The research used only open-source disinformation that was published on SEBENARNYA.MY. Neither private messages nor any personal information was gathered. When screenshots included names, usernames, profile pictures, phone numbers, or other personal data, the researchers anonymised or deleted this information during data processing. The screenshots were used only for linguistic purposes, not for personal identification. Because the research used only open-source material and no human subjects, ethical approval was not required.

Data Analysis

The analyses used qualitative textual analysis, which involved identifying and interpreting lexical features in the COVID-19 misinformation texts. A coding frame, generated from prior literature and refined during the initial analysis, was adopted to ensure each lexical feature was clearly defined before coding. All 27 misinformation texts were carefully analysed for the presence of all lexical features as defined beforehand. All occurrences were manually coded. In cases where there was overlap in categories of a given linguistic item, it was categorised depending on its primary linguistic function as described in the operational definitions. The salience of each lexical feature was based on the frequency of its occurrence and the dispersion of the features in the data set. Descriptive frequencies were computed based on the total frequency of occurrence of each feature in the 27 texts.

FINDINGS

The analysis revealed that fake news texts exhibit distinctive features such as random capitalisation patterns, including title case and all-caps words; irregular use of lowercase letters; frequent repetition of the same word; and a higher rate of incorrect spelling. Table 1 summarises the lexical features found in the fake news articles.

Table 1. Summary of Lexical Features of Fake News

Sample no.	Analysed lexical features	
	Characteristics	Example
1	Usage of random capitalisation (capitalised word)	Please Stay Away from Durian Stalls or Seller at the Roadside or Supermarkets !
2	Usage of random capitalisation (capitalised word)	Vaccination Warning
3	Usage of random capitalisation (word with all capital letters, capitalised word)	AstraZeneca Registration Early Release! APPOINTMENT BOOKINGS OPEN

		ASTRAZENECA VACCINE
4	Irregular usage of lowercase letter	...after using sinovac...
5	Usage of random capitalisation (word with all capital letters)	FINALLY DR NOOR HAS PUT HIS FOOT DOWN.
	Usage of incorrect formatting (bold)	Wednesday, February 3, 2021
6	Usage of random capitalisation (word with all capital letters, capitalised word)	BREAKING NEWS, CHINA, Causes
	Repetition of same content word	...and after examining and examining them properly,...
7	Usage of random capitalisation (word with all capital letters, capitalised word)	POLICE BULLETIN; Warning
8	Usage of random capitalisation (word with all capital letters)	...,all MUST wear face masks with effect from 1 Aug.
	Usage of incorrect formatting (bold)	Please take note.
9	Usage of random capitalisation (word with all capital letters, capitalised word)	SOME IMPORTANT POINTS TO MALAYSIANS BY THE DIRECTOR-GENERAL OF HEALTH If you Like it please Share With others ****
	Usage of incorrect formatting (bold)	THE DIRECTOR-GENERAL OF HEALTH
10	Usage of random capitalisation (word with all capital letters)	PETRONAS
	Usage of incorrect formatting (bold)	2019-nCoV : New confirmed cases in KLCC
11	Usage of random capitalisation (word with all capital letters)	So, it is IMPERATIVE that whenever u receive package from ANYONE, be it the hawker, supermarket, SANITISE/WASH your hands before you open the package. ALSO DO THE SAME AFTER opening the package.
	Irregular usage of lowercase letter	after chan death then forensic check only know he died from COVID.
12	Usage of random capitalisation (word with all capital letters)	WE WILL BE IN THE MAXIMUM STAGE OF INFECTION.

	Repetition of same content word	...as we will be at the peak of the virus is two weeks, normally in those two weeks all infected will appear then there are two weeks of clam and then two weeks where it decreases.
13	Usage of random capitalisation (word with all capital letters, capitalised word)	Do you have direct contact or is taking care of a positive COVID-19 PATIENTS? (3PTS) Do you have a travel history to a COVID-19 INFECTED AREA? Are you experiencing Fatigue?
14	Irregular usage of lowercase letter	Forwarded as received an observation of where the next likely cluster will emerge: children
15	Irregular usage of lowercase letter	a. antiviral manufacturing b. vaccine development and manufacturing c. Home respiratory support devices d. AI xray reporting for COVID 19
	Usage of incorrect spelling (scriptio continua spelling error)	
	Usage of incorrect formatting (bold)	1. Awareness Programme & Prevention
16	Usage of random capitalisation (capitalised word)	The Chief Intensivist Dato Dr Wan Nasrudin is positive for COVID 19 and is currently warded. All of you please strictly adhere to infection Control Guidelines
	Irregular usage of lowercase letter	Dr wan, Dr nazri M, Dr ruwaida, Dr rozlia, Dr mumi
	Usage of random capitalisation (word with all capital letters)	'URGENT CALL FOR SUPPORT, PLEASE VIRAL AND HELP IF YOU CAN'
	Irregular usage of lowercase letter	i thank you so much for ur help in advance
17	Irregular usage of lowercase letter	If you would like to buy the PPEs to donate hospital sungai buloh, please contact...
	Usage of incorrect spelling (non-words)	My staff are having to diy theie face shields and hoods as there are so limited amounts v for us to wear already...

		If we don't use it, we will surely get infected and wil not be able to take care of our patients...
	Repetition of same content word	We are running short of face shields, hoodsand bootcover. In fact we are running out of PPE which we use to cover ourselves...
	Usage of random capitalisation (capitalised word)	Dear All, just received the latest press release from the police: The Police will ...
18	Repetition of same content word	They are asking out now to go back home immediately. During the road block and anyone caught not wearing mask will be fine Rm200 immediately.
19	Usage of random capitalisation (word with all capital letters, capitalised word)	COVID PATIENT!! This Chinese lady is from Simee, Ipoh, just escaped from Ipoh General Hospital, wanted by the Hospital and Police.
20	Usage of random capitalisation (capitalised word)	Starting today – Drive Through (COVID-19 Testing) Klinik Kesihatan Tanglin, at Lake Garden KL. Free Consultation.
	Usage of incorrect spelling (addition)	Informations
21	Usage of random capitalisation (word with all capital letters)	COVID 19 ALERT FROM DOSH IN PASIR GUDANG
22	Irregular usage of lowercase letter	One of his patient caught the corona virus from Japanese pt who was admitted in Pantai – sanjiv's pt was her roommate.
	Usage of random capitalisation (word with all capital letters, capitalised word)	Confirm News ! This stupid women and husband did not isolated themselves after back from ITALY.
23	Usage of incorrect formatting (word spacing)	Myfriend gotvirus
	Usage of incorrect spelling (scriptio continua error)	Myfriend coming back from ITALY and his colleagues gotvirus at the aeroplane.

24	Usage of random capitalisation (capitalised word)	Kindly contact the nearest Hospital or Clinic.
25	Usage of random capitalisation (word with all capital letter, capitalised word)	*Penang Already have Corona* Loh Guan Lye already confirmed a CHINA Guest have one Corona Virus Positive cases
26	Irregular usage of lowercase letter	1.repeated high fever 2.prolonged coughing after fever 3.Children are prone 4. Adults usually feel uneasy, headache and mainly 5.highly contagious
	Usage of incorrect formatting (bold)	Ministry of health's notification to the public that the influenza this time is serious.
27	Usage of random capitalisation (capitalised word)	The patient was confirmed dead due to COVID in the ICU ward by a Malay Lady Doctor on call. The Ambulance with Doctor and Nurses immediately arrived the burial site and examined him and found that he was still alive.

Usage of Random Capitalisation

English capitalisation rules govern the use of uppercase letters at the start of a word to signal the beginning of a sentence or quotation, proper nouns, and adjectives derived from proper nouns. Capitalisation is also used for the first-person singular pronoun, acronyms (Pathan, 2021), honorary and professional titles, familial relationships used as titles, and newspaper datelines. However, the majority of the fake news articles (22 out of 27) used capitalisation randomly, not following the guidelines of English capitalisation. The capitalisation patterns observed in the data are categorised as 'random capitalisation', which can be subdivided into two: capitalised words or title case, and words in all capital letters. These categories are described in detail below.

Usage of Capitalised Word/Title Case

In this analysis, the initial letter of every word in a sentence was randomly capitalised. This inappropriate use of title case has been observed in fake news articles, as seen in the following sentence sample no.1 (Figure 2): "Please **Stay Away** from Durian **Stalls** or **Seller** at the **Roadside** or **Supermarkets** !"

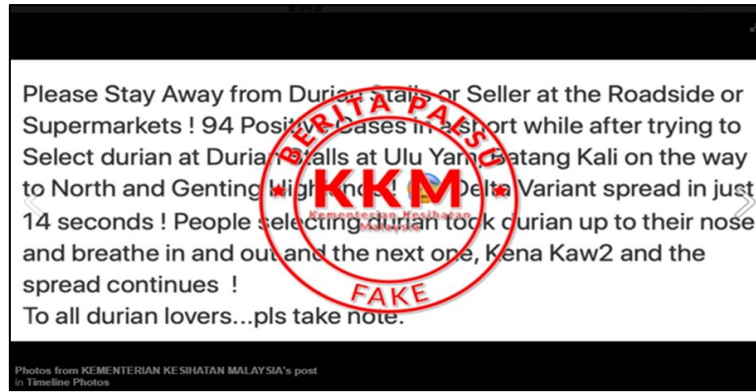


Figure 2. Usage of title case in sentences from a Malaysian fake news text.

The example above shows that capitalised words appear across word classes, except for the preposition (from, at), the conjunction (or), and the definite article (the). Only content words such as proper nouns (Durian Stalls), common nouns (Seller, Roadside, Supermarkets), verb (Stay Away), and adverb (Please) are capitalised.

Usage of Words in all Capital Letters

This category showcases the words written in all capital letters or uppercase. Generally, this type of capitalisation is used in newspaper dateline in informing readers on the source of the news and mentioning certain organisations such as WHO (World Health Organization) and UNICEF (United Nations Children's Fund). It is primarily deduced that capitalisation is most likely to draw audiences' attention to the main intended message in the articles. The data analysis revealed that words written entirely in capital letters were widely used across all word classes in the fake news texts, serving different linguistic functions. For example, there are instances in which a whole sentence was written in all uppercase letters irrespective of their word classes, as evident in the following sample no. 12 (Figure 3): “*DO NOT STAY WITH THIS MESSAGE, PASS IT ON TO ALL YOUR CONTACTS*”.

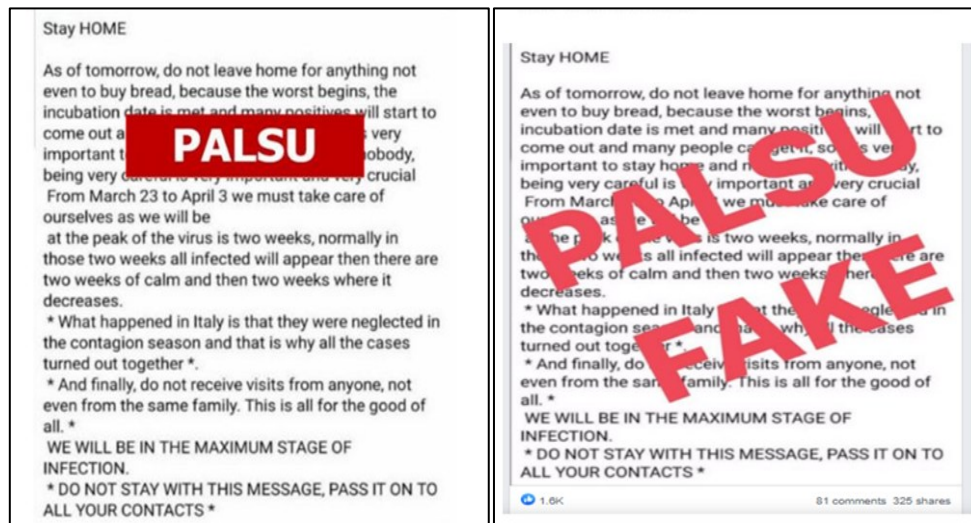


Figure 3. Example of all-capital-letter word usage as an orthographic feature employed for emphasis in a fake news article.

Further analysis categorised the grammatical classes of words written entirely in capital letters as they appear in the data sample. One such analysis is illustrated below in Figure 4, wherein the words written in capitals belong to different word classes such as verbs (DO, PASS), adverbs (NOT), nouns (MESSAGE, CONTACTS), pronouns (IT, YOUR), determiners (THIS, ALL), prepositions (WITH, ON), and the infinitive marker (TO). This analysis shows that capital letters are not confined to a particular word class but are used across various classes to create a sense of urgency.

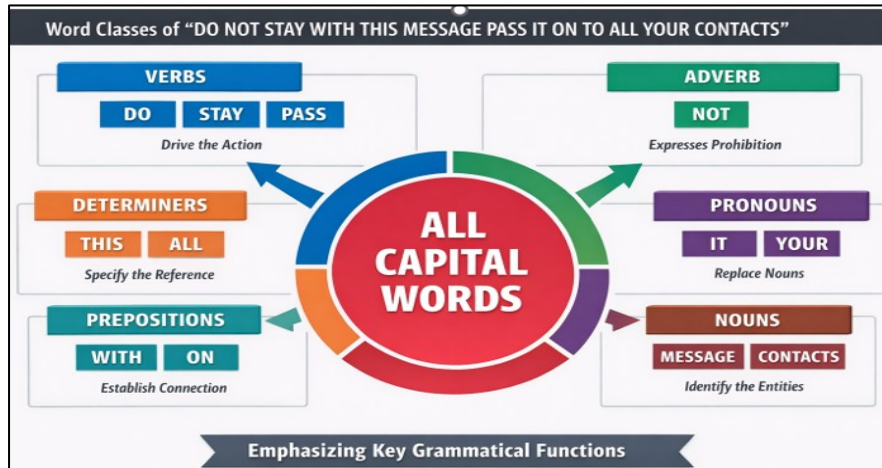


Figure 4. Parts of speech of the capitalised words in the message

Irregular Usage of Lowercase Letter

The data analysis identified another distinctive lexical feature, which is the irregular use of lowercase letters. Several instances were observed in which lowercase letters are used in violation of standard English grammatical and orthographic conventions, making it a significant lexical marker of fake news. An example of this was found in somewhat genuine-looking content that provided a list of required technologies to address the COVID-19 epidemic (Refer to Figure 5). Despite the first positive impression of the article which bears the official Malaysian coat of arms, closer inspection on the text found that the capitalization format of the items listing was not consistent, where some of the subitems listed would start with capital letters and the others would start with lowercase letters. The first instance of this feature can be observed in Sample No. 15 (Figure 5), where under item '4. Treatment', the initial letters of the first two subitems, (a) and (b) are in lowercase letters, whereas the following two items, (c) and (d), begin with uppercase letters:

- (a) Antiviral manufacturing
- (b) Vaccine development and manufacturing
- (c) Home respiratory support devices
- (d) AI X-ray reporting for COVID19

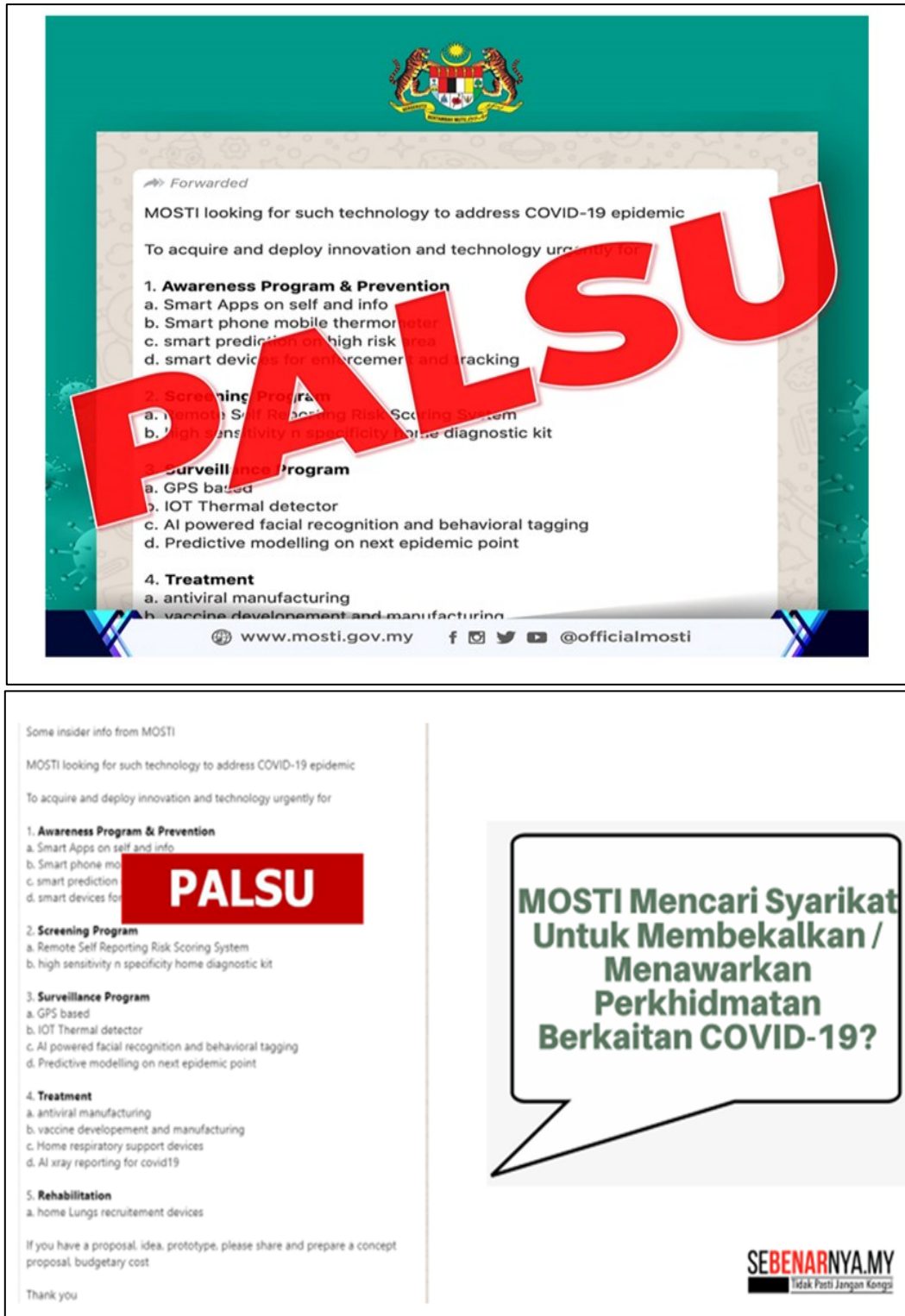


Figure 5. Examples of irregular lowercase letter usage.

The irregularity in formatting within a single text sample is unusual for supposedly official government content or a professionally written text addressing a serious global health crisis. The same feature has also been identified in text sample no. 26 (refer to Table 1). The analysis also shows that fake news uses lowercase first letters in proper nouns (e.g., names of an individual/place), as evident in these items from sample no. 16 (Figure 6): ‘Dr wan’, ‘Dr nazri M’, ‘Dr ruwaida’, ‘Dr rozlia’, and ‘Dr mumi’.

The same feature is also evident in Text Sample 26 (see Table 1). In this message, irregular lowercase usage occurs in the phrase "Ministry of Health's notification to the public that the influenza this time is serious," where the common noun "health" is incorrectly written in lowercase as part of the official institutional name "Ministry of Health." According to English spelling conventions, all principal words in the names of government agencies must be capitalised. This inconsistency violates writing conventions. Moreover, extra bold styles are used in expressions such as "repeated high fever," "prolonged coughing after fever," "Children are prone," "Adults usually feel uneasy, headache and mainly," and "highly contagious." This makes these phrases more noticeable to the reader and highlights important statements related to health issues. Such orthographic and formatting irregularities are characteristic of informal online misinformation messages, where visual emphasis is often employed to enhance urgency and persuade readers of the message's credibility.

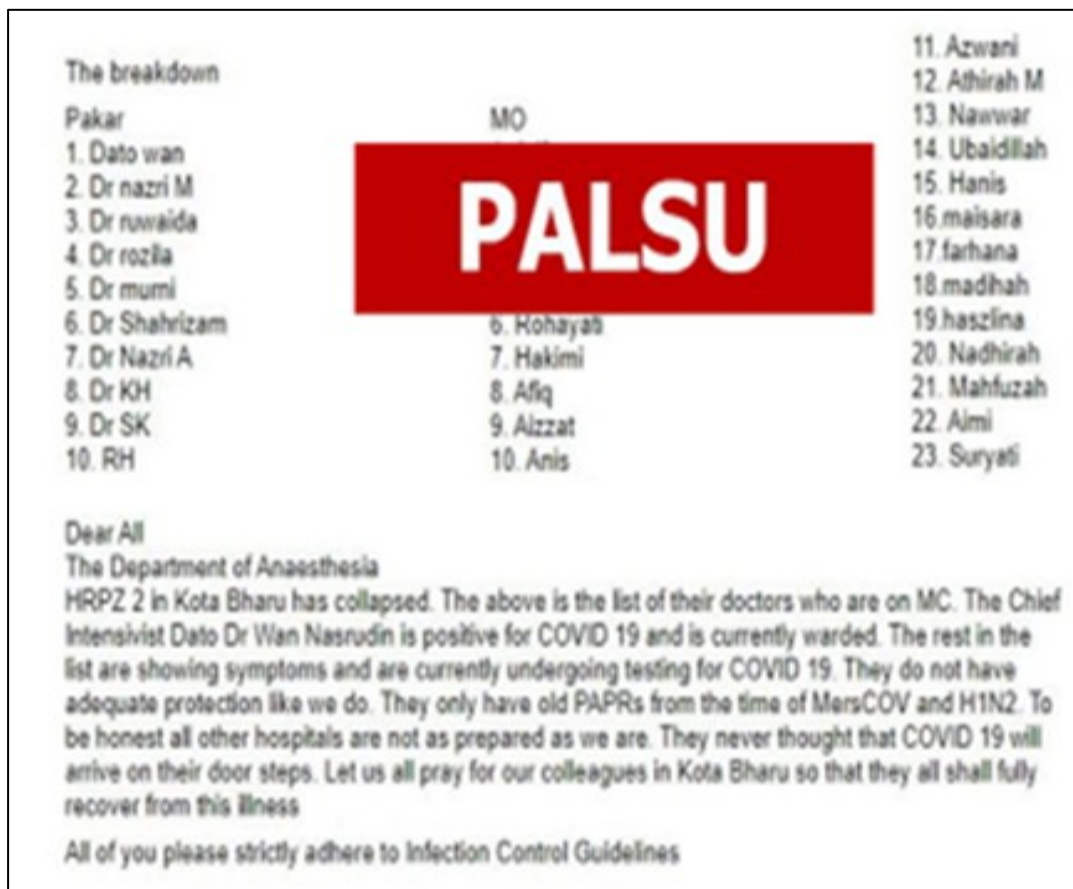
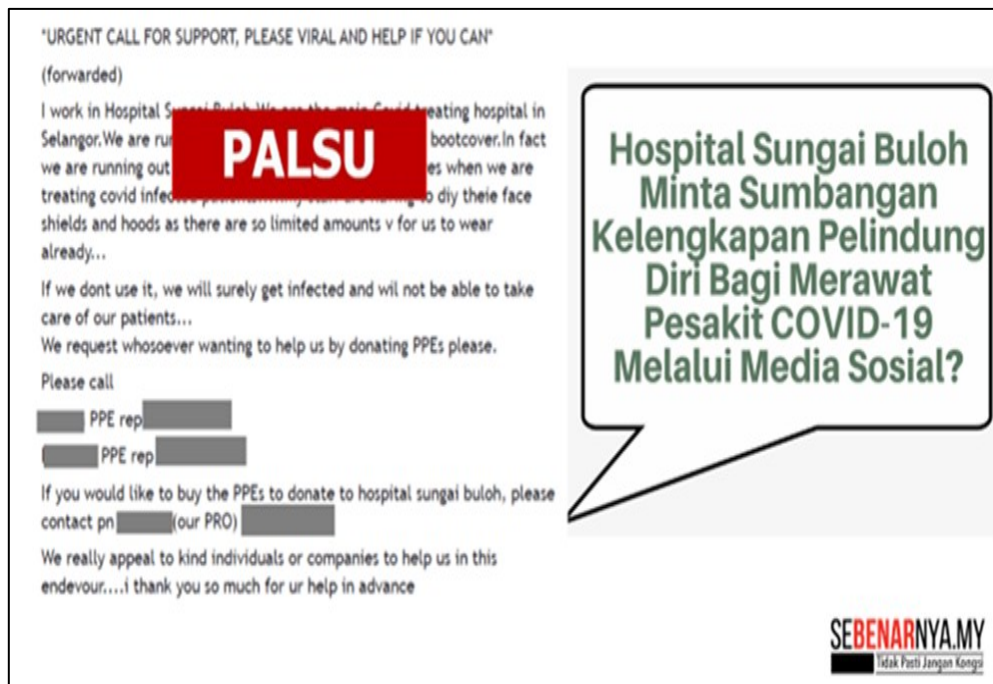


Figure 6. Example of improper lowercase usage in the initials of proper nouns (names of individuals)

In an example from sample no. 17 (Figure 7), it was also found that the proper noun 'hospital sungai buloh' which is the name of a government hospital, is written in lowercase: "If you would like to buy the PPEs to donate **hospital sungai buloh**, please contact...". From the same sample, it was observed that a similar pattern emerges where the first-person singular pronoun 'I' is written in lowercase letter 'i' instead of the uppercase letter 'I' as seen in this sentence: "i thank you so much for ur help in advance". From these examples, it is inferred that the creators of fake news are either unaware of proper writing following English grammatical rules, or have neglected to rectify writing mistakes in their content. This may also signal that the news is not from legitimate sources, as the articles were written while

disregarding formal writing rules.



(a) Transcript of fake message



(b) Original text from messaging platform

Figure 7. Example of multiple lexical markers of fake news, including irregular lowercase usage, repetition of content words, and incorrect spellings

Repetition of the Same Content Words

Previous studies have found that fake news tends to use repetitive language (Horne & Adali, 2017) and shows less lexical variation than real news. Creators of fake news are mostly thought to intentionally create false narratives, which factually requires greater cognitive ability than directly reporting real news. However, due to the high cognitive demand of fabricating stories, the language of fake news could become much simpler instead (Newman et al., 2003) and a lot of repetitive items may be observed in fake news content. The following sentence in sample no. 12 (Figure 3) demonstrates this fact: "...as we will be at the peak of the virus is **two weeks**, normally in those **two weeks** all infected will appear then there are **two weeks** of calm and then **two weeks** where it decreases.". In this short sample alone, the word 'two weeks' have been repeated as many as four times. Notably, only content words that were found to be used repetitively in the data analysis.

Usage of Incorrect Spelling

Another significant lexical marker to detect fake news is incorrect spelling. There are four types of spelling errors identified in the data, such as addition, non-words, omission, and scriptio continua spelling error. This lexical feature was found in sample no.17 (Figure 7) as shown in these sentences: "My staff are having to **diy theie** face shields and hoods as there are so limited amounts v for us to wear already..."; and "If we don't use it, we will surely get infected and **wil** not be able to take care of our patients...". From this sample alone three misspelled items were identified. The word '**wil**' is a spelling error by omission of the letter 'l' at the end of the word, meanwhile the other two mistakes, '**diy**' and '**theie**' were identified as non-words. A non-word spelling error might occur due to the writer being unaware of the correct spelling, while omission can either be a typographical error or also the writer's insufficient knowledge on the correct spelling.

This study identified a distinctive spelling pattern called a scriptio continuum spelling error, as illustrated in Sample 23 (Figure 8): "Myfriend coming back from ITALY and his colleagues gotvirus at the aeroplane." In the above-mentioned example, the combination of "Myfriend" and "gotvirus" can be explained by the fact that the necessary spaces separating 'My' from 'friend' and 'got' from 'virus' were missing. As a result, non-standard word formations emerged, which differed from standard English spelling norms. This phenomenon is similar to Scriptio Continua, an ancient writing style in which words were written without separation. However, in this research, such phenomena are considered spelling mistakes.



Figure 8. Example of scriptio continua spelling error and incorrect word spacing

Usage of Incorrect Formatting

Improper formatting concerns the wrong presentation of text, not incorrect spelling. While scriptio continua errors refer to spelling mistakes that do not conform to the existing spelling rules, formatting mistakes are done intentionally to highlight specific points. In Sample 26 (Table 1), there is the use of bold text for terms like “repeated high fever,” “coughing after fever,” and “highly contagious.” These formatting choices enhance the perceived urgency and salience of the message and may increase its persuasive impact, even though they do not affect the spelling of individual words. It was also found that fake news uses bold text to emphasise the message, although this is not standard writing format. An example is seen in a sentence from sample no. 26 (Figure 9) as follows: “**Ministry of health’s notification to the public that the influenza this time is serious.**”.

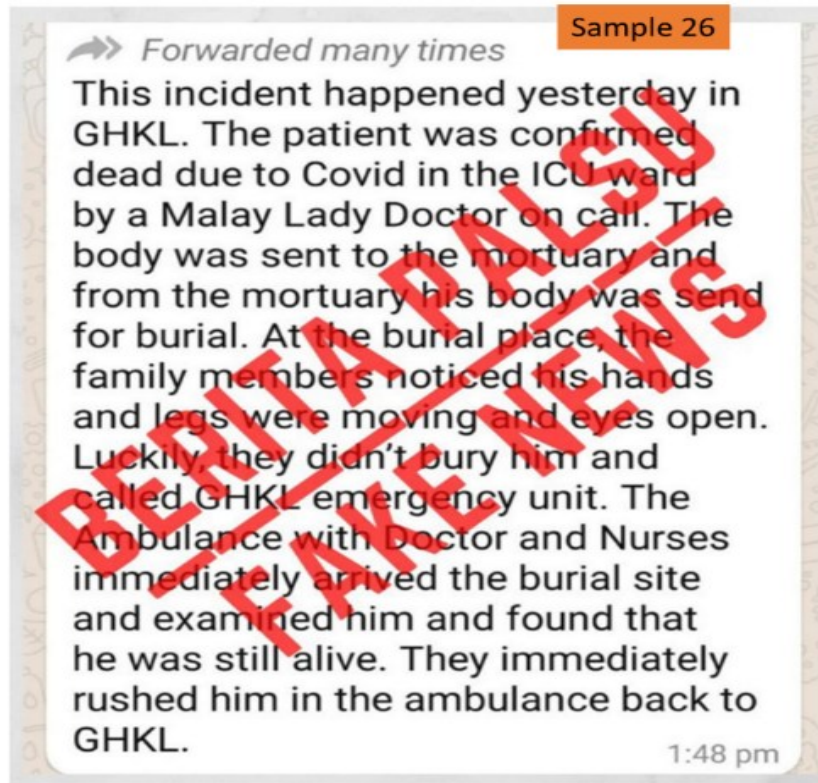


Figure 9. Example of scriptio continua spelling error and incorrect word spacing, where multiple words are written without spaces

Table 2. Frequency Distribution of Lexical Features in the COVID-19 Misinformation Corpus.

Lexical Feature	Subcategory	No. of Texts (n = 27)	Percentage (%)
Random capitalisation	Capitalised words and/or ALL CAPS	22	81.5
Irregular lowercase letters	Non-standard use of lowercase	7	25.9
Repetition	Repetition of content words or phrases	4	14.8
Spelling errors	Scriptio continua	2	7.4
	Non-word spelling	1	3.7
	Addition error	1	3.7
Total spelling errors	All spelling error types combined	4	14.8
Incorrect formatting	Excessive bold text	5	18.5
	Improper word spacing	1	3.7
Total incorrect formatting	All formatting features combined	6	22.2

Table 2 presents the distribution of lexical features identified across the 27 COVID-19 misinformation texts. Overall, random capitalisation was the most prevalent lexical feature, occurring in 22 of the 27 texts (81.5%), indicating that unconventional use of capital letters was a dominant characteristic of the corpus. This included both unnecessary initial capitalisation and words written entirely in uppercase to emphasise particular information. The second most frequently identified feature was incorrect formatting, which appeared in 6 texts (22.2%). Most formatting irregularities involved the excessive use of bold text (5 texts; 18.5%), while improper word spacing was identified in one text (3.7%). Irregular use of lowercase letters was observed in 7 texts (25.9%), reflecting deviations from standard English capitalisation conventions.

Less frequently occurring features included repetition of content words or phrases and spelling errors, each identified in 4 texts (14.8%). Among the spelling-error subcategories, scriptio continua were the most common, occurring in 2 texts (7.4%), whereas non-word spelling and addition errors each appeared in one text (3.7%).

In summary, the results suggest that orthographic features, especially random capitalisation, occurred more frequently than lexical features such as repetition and misspellings within the analysed corpus of misinformation. These frequencies give a general idea of lexical feature occurrence and highlight the most notable ones for further discussion.

DISCUSSION

The current study intends to examine the lexical characteristics that define fake COVID-19 news and determine which of those features can be considered the most effective way of differentiating fake news from real ones. Drawing on forensic linguistics, linguistic feature analysis, and the Unnddeutsch Hypothesis, the results reveal a distinctive set of lexical irregularities in fake news during the COVID-19 pandemic. Rather than chance misprints and errors, these peculiarities serve as persuasive, emotive, and attention-grabbing means that contribute to the deceptive nature of fake news.

In contrast to previous research, the present study validates and extends the prior body of knowledge by emphasising the importance of human-traceable surface lexical features rather than algorithmic ones. One of the most significant findings of the current study is the abundance of random capitalisation in fake news, i.e. both capitalised words and full uppercase sentences. This is consistent with previous research showing that fake news tends to use typography to control readers' attention and create an impression of urgency and authority (Horne & Adali, 2017; Zhou, 2021). While real news follows institutional editorial standards and orthographic rules, the visual presentation of fake news is inconsistent, reflecting a lack of professionalism and effective gatekeeping.

Previous corpus-based studies show that sensationalism in fake news is expressed through exaggerated punctuation, emotive adjectives, and formatting techniques such as excessive capitalisation (Antypas et al., 2021; Qi et al., 2023). The current study goes beyond previous research by showing that capitalisation in fake news is not limited to headlines or key nouns but appears across various grammatical classes, including verbs, pronouns, determiners, and prepositions. It disrupts grammatical hierarchy and suggests a deliberate effort to make a message visually visible rather than structurally coherent. From a forensic linguistics perspective, such orthographic inconsistency can be considered a trace of deception, where persuasion prevails over clarity and truth (Sousa-Silva, 2022). Moreover, the presence of full uppercase sentences, especially imperative ones like "DO NOT STAY WITH THIS

MESSAGE, PASS IT ON," proves a communication strategy of making a message go viral. This aligns with Bruns et al. (2022), who assume that fake news messages include calls to action to share quickly and spread misinformation before it is verified.

Another important lexical feature is irregular lowercase usage, namely in proper nouns, sentence beginnings, and the first-person singular pronoun "I." Those deviations from English orthography norms are rare in real news, which passes strict editorial checks. It is consistent with the research by Mahyoob et al. (2021), which states that fake news is characterised by poor grammatical accuracy and low adherence to the rules of formal writing. In terms of the Unndeutsch Hypothesis, those irregularities can be seen as cognitive artefacts of fabrication. Invented narratives consume much more cognitive capacity because authors must invent details rather than recall them, and they often receive less attention to linguistic form (Newman et al., 2003). Current research confirms this, showing that fake news authors prioritise message dissemination over linguistic precision. In contrast, real news based on verifiable events is structurally cohesive and orthographically consistent (Grieve & Woodfield, 2023). Notably, the misuse of lowercase letters in professional titles, people's names, and references to institutions (e.g., "hospital sungai buloh," "Dr wan") reduces the text's credibility while trying to project authority. This contradiction reveals a fundamental conflict between fake news and real journalism: a need to appear authoritative without the linguistic competence and infrastructure of real journalism.

The current analysis shows extensive repetition of the same words, phrases, and clauses in fake news texts. This pattern is well known in deception research. According to Newman et al. (2003), deceptive texts tend to use repetitive, simplistic language because deception imposes a high cognitive load. Similarly, Antypas et al. (2021) found that COVID-19 misinformation tweets exhibit less lexical variety and a narrower topical range than truthful content. The current research confirms these findings and shows that repetitions in fake news texts are usually strategically arranged. It aligns with Wang et al. (2022), who found that fake COVID-19 narratives often use reiteration to strengthen ideologically framed causal relationships. Real news, by contrast, usually avoids extensive repetition and prefers lexical variety and informational density to maintain clarity and readers' interest (Ghayoomi, 2022). Thus, repetition becomes a strong lexical sign of deception, perceptible to readers without special knowledge or technology.

Spelling errors become another prominent lexical characteristic of fake news texts. They can vary from omission, addition, homophonic substitutions, non-words, incorrect kerning, to scriptio continua. This aligns with observations by Horne and Adalı (2017), who found that fake news contains more typographical and orthographic mistakes than real news. However, the discovery of scriptio continua spelling errors is the original contribution of the present research. The absence of word boundaries (e.g., "myfriend," "gotvirus") is particularly striking, since such forms are hard to tolerate in edited texts. From the perspective of forensic linguistics, those anomalies indicate a spontaneous or hurried creation of the text aimed to quick distribution rather than editing. This aligns with Zhang and Ghorbani (2020), who argue that fake news emphasises speed and virality over accuracy and linguistic precision. Spelling errors involving technical or scientific terminology (e.g., "PH" instead of "pH," "CRONA" instead of "CORONA") are especially informative in COVID-19 misinformation. Undoubtedly, fake news often adopts scientific discourse without understanding its rules, resulting in superficial imitation rather than competence. Unlike many previous studies focused on machine-learning classifiers (Antypas et al., 2021; Qi et al., 2023), this research emphasises human-recognisable lexical features.

Detecting human-recognisable lexical features is indeed a crucial contribution, as algorithmic approaches often remove precisely those features, such as capitalisation, spelling, and formatting, that are most informative for human judgement. By identifying surface-level lexical indicators, this study empowers readers to detect fake news in real time, without reliance on automated systems or delayed fact-checking mechanisms. In line with Koch et al. (2023), this study suggests that strengthening public awareness of linguistic deception strategies may be more effective than reactive platform interventions alone. From a socio-psychological perspective, recognising these lexical patterns can reduce susceptibility to emotionally manipulative content and enhance critical digital literacy. Overall, the findings strongly support the Unnddeutsch Hypothesis, demonstrating that fabricated COVID-19 news systematically diverges from authentic reporting at the lexical level. Fake news texts are characterised by orthographic instability, repetitive vocabulary, spelling irregularities, register mismatch, and exaggerated visual emphasis. Compared with existing literature, this study not only confirms earlier observations but also extends them through fine-grained qualitative analysis of Malaysian COVID-19 fake news discourse. By positioning lexical features as accessible and interpretable markers of deception, this research contributes to linguistic fake news studies and offers practical implications for education, media literacy, and public health communication.

CONCLUSION

This article combines forensic linguistic analysis with linguistic analysis of fake news to understand the linguistic resources of fake news and identify potential indicators. The current study found that random capitalisation, repetition of the same content word, and different types of spelling mistakes are the most significant indicators of fake news. The other potential lexical markers of fake news include usage of incorrect formatting, informal shortened words, and inappropriate word choice. These findings are very promising because they may help people not only spot fake news in real-time without any fact-checking tools but also improve algorithms for AI to identify fake news.

REFERENCES

- Ali, A. A., Latif, S., Ghauri, S. A., Song, O. Y., Abbasi, A. A., & Malik, A. J. (2023). Linguistic Features and Bi-LSTM for Identification of Fake News. *Electronics*, 12(13), 2942. <https://doi.org/10.3390/electronics12132942>
- Ali, M., Liu, H., & Zhai, X. (2023). Fake news detection on social media: A systematic review. *Information Processing & Management*, 60(1), 102878. <https://doi.org/10.1016/j.ipm.2022.102878>
- Antypas, D., Preece, A., & Taylor, I. (2021). Detecting COVID-19 misinformation on Twitter using lexical features. *Applied Soft Computing*, 104, 107224. <https://doi.org/10.1016/j.asoc.2021.107224>
- Baxter, J. (2020). International Encyclopedia of Human Geography (2nd edition). *Text and Textual Analysis*, 13, 239–243. <https://doi.org/10.1016/B978-0-08-102295-5.10872-8>
- Biber, D., & Conrad, S. (2019). Register, genre, and style (2nd ed.). *Cambridge University Press*. <https://doi.org/10.1017/9781108686136>
- Bruns, A., Harrington, S., & Hurcombe, E. (2022). ‘Corona? 5G? Or both?’: The dynamics of COVID-19 conspiracy theories on social media. *Media International Australia*, 177(1), 12–29. <https://doi.org/10.1177/1329878X20946113>

- Ghayoomi, M., & Mousavian, M. (2022). Deep transfer learning for COVID-19 fake news detection in Persian. *Expert Systems*, 39(8), e13008. <https://doi.org/10.1111/exsy.13008>
- Grieve, J., & Woodfield, H. (2023). Linguistic evidence of fabrication in journalistic writing. *Applied Linguistics*, 44(2), 259–282. <https://doi.org/10.1093/applin/amac021>
- Grieve, J., & Woodfield, H. (2023). *The Language of Fake News*. Cambridge University Press. <https://doi.org/10.1017/9781009349161>
- Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 759–766. <https://doi.org/10.1609/icwsm.v11i1.14976>
- Koch, T. K., Frischlich, L., & Lerner, E. (2023). Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media. *Journal of Applied Social Psychology*, 53(6), 495–507. <https://doi.org/10.1111/jasp.12959>
- Kondamudi, M. R., Sahoo, S. R., Chouhan, L., & Yadav, N. (2023). A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. *Journal of King Saud University-Computer and Information Sciences*, 35(6), 101571. <https://doi.org/10.1016/j.jksuci.2023.101571>
- Mahyoob, M., Elyas, T., & Alqarni, N. (2021). Linguistic features of fake news in social media. *Cogent Arts & Humanities*, 8(1), 1879086. <https://doi.org/10.1080/23311983.2021.1879086>
- Wang, Y., Ma, F., Jin, Z., et al. (2022). Does fake news in different languages tell the same story? An analysis of multi-level thematic and emotional characteristics of news about COVID-19. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-022-10329-7>
- Zhou, C., Li, K., & Lu, Y. (2021). Linguistic characteristics and the dissemination of misinformation in social media: The moderating effect of information richness. *Information Processing & Management*, 58(6), 102679. <https://doi.org/10.1016/j.ipm.2021.102679>
- Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5), 665–675. <https://doi.org/10.1177/0146167203029005010>
- Pathan, M. A. K. (2021). The most frequent capitalization errors made by the EFL learners at undergraduate level: an investigation. *Scholars International Journal of Linguistics and Literature*, 4(3), 65-72. <https://doi.org/10.36348/sijll.2021.v04i03.001>
- Petrou, N., Christodoulou, C., Anastasiou, A., Pallis, G., & Dikaiakos, M. D. (2023). A Multiple change-point detection framework on linguistic characteristics of real versus fake news articles. *Scientific Reports*, 13(1), 6086. <https://doi.org/10.1038/s41598-023-32952-3>
- Rastogi, S., & Bansal, D. (2023). A review on fake news detection 3T's: Typology, time of detection, taxonomies. *International Journal of Information Security*, 22(1), 177-212. <https://doi.org/10.1007/s10207-022-00625-3>

- Shahzad, K., Khan, S. A., Iqbal, A., Shabbir, O., & Latif, M. (2023). Determinants of fake news diffusion on social media: a systematic literature review. *Global Knowledge, Memory and Communication*. <https://doi.org/10.1108/GKMC-06-2023-0189>
- Sousa-Silva, R. (2022). Fighting the fake: A forensic linguistic analysis to fake news detection. *International Journal for the Semiotics of Law-Revue*, 35(6), 2409-2433. <https://doi.org/10.1007/s11196-022-09901-w>
- Tiersma, P. M., & Solan, L. M. (2012). The language of crime. In P. M. Tiersma & L. M. Solan (Eds.), *The Oxford handbook of language and law* (pp. 340–353). *Oxford University Press*. <https://doi.org/10.1093/oxfordhb/9780199572120.013.0025>
- Trnavac, R., & Pöldvere, N. (2024). Investigating appraisal and the language of evaluation in fake news corpora. *Corpus Pragmatics*, 8(2), 107–130. <https://doi.org/10.1007/s41701-023-00162-x>
- Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2), 102025. <https://doi.org/10.1016/j.ipm.2019.03.004>