eISSN : 2287-4577
pISSN : 2287-9099
https://www.jistap.org
Vol. 13 No. 1 March 30, 2025

Journal of Information Science Theory and Practice

Journal of Information Science Theory and Practice

JISTaP

01 Do Retweets Undermine Twitter Mentions of Journal Articles as Indicators of Research Significance?
15 Topic Modeling Approach on Twitter Data Relevant to Political Changes During the COVID-19 Pandemic
36 Library Space and Resource Usage Time: Reflections from Digital Footprints During Pre and Post-COVID-19
51 Hate Speech Identification and Categorization on Social Media Using Bi-LSTM: An Information Science Perspective
70 How Do Open and Closed Access Journals Compare in Citations, Altmetrics, and Social Media Engagement for Pesticide Research?
85 Enhancing QA System Evaluation: An In-Depth Analysis of Metrics and Model-Specific Behaviors

Vol. 13 I No. 1 I March 30, 2025

KISTI www.kisti.re.kr

Indexed/Covered by SCOPUS, LISA, DOAJ, CrossRef, and KCI

JISTaP
Journal of Information Science Theory and Practice

General Information

Aims and Scope

The Journal of Information Science Theory and Practice (*JISTaP*) is an international journal that aims at publishing original studies, review papers and brief communications on information science theory and practice. The journal provides an international forum for practical as well as theoretical research in the interdisciplinary areas of information science, such as information processing and management, knowledge organization, scholarly communication and bibliometrics. JISTaP will be published quarterly, issued on the 30th of March, June, September, and December. JISTaP is indexed in the Scopus, Korea Citation Index (KCI), Directory of Open Access Journal (DOAJ), Korea Science Citation Index (KSCI) and KoreaScience by the Korea Institute of Science and Technology Information (KISTI) as well as CrossRef. The full text of this journal is available on the website at http://www.jistap.org

Indexed/Covered by

Scopus LISA DOAJ DIRECTORY OF OPEN ACCESS JOURNALS Crossref Cited-by 한국학술지인용색인 Korea Citation Index

Publisher

Korea Institute of Science and Technology Information
66, Hoegi-ro, Dongdaemun-gu, Seoul, Republic of Korea
(T) +82-2-3299-6105
(F) +82-2-3299-6067
E-mail: jistap@kisti.re.kr
URL: http://www.jistap.org

Managing Editor: Hyekyong Hwang, Eungi Kim

Copy Editor: Ken Eckert

Design & Printing Company: MEDrang Inc.

2nd Floor, Hyoryeong Building, 32 Mugyo-ro, Jung-gu, Seoul 04521, Korea
(T) +82-2-325-2093
(F) +82-2-325-2095
E-mail: info@medrang.co.kr
Homepage: www.medrang.co.kr

Open Access and Creative Commons License Statement

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. Open Access articles are automatically archived in the Korea Institute of Science and Technology Information (KISTI)'s Open Access repository (AccessON). All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/). Under this license, the authors retain full ownership of their work, while permitting anyone to use, distribute, and reproduce the content in any medium, as long as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

This paper meets the requirements of KS X ISO 9706, ISO 9706-1994 and ANSI/NISO Z39.48-1992 (Permanence of Paper).

2025 Copyright © Korea Institute of Science and Technology Information

JISTaP
Journal of Information Science Theory and Practice

Editorial Board

Co-Editors-in-Chief

Gary Marchionini
University of North Carolina, USA

Dong-Geun Oh
Keimyung University, Korea

Co-Associate Editors

Kiduk Yang
Youngnam Old Document Archive Center, Korea

Hyejin Lee
Korea Institute of Science and Technology Information, Korea

Co-Managing Editors

Hyekyong Hwang
Korea Institute of Science and Technology Information, Korea

Eungi Kim
Keimyung University, Korea

Editorial Board

Dan Albertson
University at Buffalo, SUNY, USA

Daniel Martínez Ávila
San Paulo State University, Brazil

Beeraka Ramesh Babu
University of Madras, India

Pia Borlund
University of Copenhagen, Denmark

France Bouthillier
McGill University, Canada

Kathleen Burnett
Florida State University, USA

Ina Fourie
University of Pretoria, South Africa

Boryung Ju
Louisiana State University, USA

Thomas Mandl
Universiät Hildesheim, Germany

Lokman I. Meho
American University of Beirut, Lebanon

Jin Cheon Na
Nanyang Technological University, Singapore

Dan O'Connor
Rutgers University, USA

M. P. Satija
Guru Nanak Dev University, India

Ou Shiyan
Nanjing University, China

Paul Solomon
University of South Carolina, USA

Kulthida Tuamsuk
Khon Kaen University, Thailand

Lihong Zhou
Wuhan University, China

Xiao-Zhi Gao
University of Eastern Finland, Finland

Wole Michael Olatokun
University of Ibadan, Nigeria

Consulting Editors

Reysa Alenzuela
Oriental Institute, ASCR, Czech Republic

Wayne Buente
University of Hawaii, USA

Sujin Butdisuwan
Mahasarakham University, Thailand

Seon Heui Choi
Korea Institute of Science and Technology Information, Korea

M. Krishnamurthy
DRTC, Indian Statistical Institute, India

S.K. Asok Kumar
The Tamil Nadu Dr Ambedkar Law University, India

P. Rajendran
SRM University, India

2025 Copyright © Korea Institute of Science and Technology Information

JISTaP
Journal of Information Science Theory and Practice

Table of Contents

Vol. 13 No. 1 March 30, 2025
Journal of Information Science Theory and Practice • https://www.jistap.org

Articles

01 Do Retweets Undermine Twitter Mentions of Journal Articles as Indicators of Research Significance?
Tint Hla Hla Htoo, Na Jin-Cheon, Michael Thewall

15 Topic Modeling Approach on Twitter Data Relevant to Political Changes During the COVID-19 Pandemic
M. G. D. S. Hansika, K. S. Ranasinghe, R. A. H. M. Rupasingha

36 Library Space and Resource Usage Time: Reflections from Digital Footprints During Pre and Post-COVID-19
Wannisa Matcha, Nawapon Kaewsuwan, Komgrit Rumdon

51 Hate Speech Identification and Categorization on Social Media Using Bi-LSTM: An Information Science Perspective
Krishna Kumar Mohbey, Basant Agarwal, Nishtha Kesswani, Maxim Sterjanov, Yunevich Nikol, Vishnyakova Margarita

70 How Do Open and Closed Access Journals Compare in Citations, Altmetrics, and Social Media Engagement for Pesticide Research?
Sivagami Murugappan, Jeyshankar Ramalingam

85 Enhancing QA System Evaluation: An In-Depth Analysis of Metrics and Model-Specific Behaviors
Heesop Kim, Aluko Ademola

99 **Call for Paper**

100 **Information for Authors**

2025 Copyright © Korea Institute of Science and Technology Information

JISTaP https://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 13(1): 1-14, 2025

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2025.13.1.1

Do Retweets Undermine Twitter Mentions of Journal Articles as Indicators of Research Significance?

Tint Hla Hla Htoo*
Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore
E-mail: tint0010@e.ntu.edu.sg

Na Jin-Cheon
Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore
E-mail: TJCNa@ntu.edu.sg

Michael Thewall
Information School, University of Sheffield, Sheffield, United Kingdom
E-mail: m.a.thelwall@sheffield.ac.uk

ABSTRACT

This study aims to investigate whether the prevalence of retweets challenges the validity of Twitter mentions (tweet counts and tweeter counts) as indicators of research significance. The study first examines whether the tweeting of research papers signifies either the wisdom of crowd effect or herd-like behavior for a dataset of COVID-19 papers from *The Lancet*. The study then uses the Modality, Agency, Interactivity, and Navigability (MAIN) model to examine the nature of the influence involved in retweeting. The Mann-Whitney U test and multiple linear regression were used. Findings show that there was extensive evidence of herd-like behaviour, rather than the wisdom of crowd effect, in tweeting research papers, challenging the validity of Twitter mentions as indicators of research significance. Credibility heuristic cues of the original tweeters were associated with their retweet rates, suggesting that retweets are more likely influenced by perceptions of the original tweeter's credibility rather than the quality of research papers.

Keywords: altmetrics, Twitter mentions, MAIN model, heuristic cues

Received: January 14, 2024 **Revised:** August 15, 2024
Accepted: August 16, 2024 **Published:** March 30, 2025

***Corresponding Author:** Tint Hla Hla Htoo
https://orcid.org/0000-0001-9331-5260
E-mail: tint0010@e.ntu.edu.sg

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. Open Access articles are automatically archived in the Korea Institute of Science and Technology Information (KISTI)'s Open Access repository (AccessON). All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/). Under this license, the authors retain full ownership of their work, while permitting anyone to use, distribute, and reproduce the content in any medium, as long as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© 2025 [Tint Hla Hla Htoo, Na Jin-Cheon, Michael Thewall] This article is distributed under the Creative Commons Attribution License (CC BY), allowing unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. It is published by the Korea Institute of Science and Technology Information (KISTI).

1. INTRODUCTION

As part of the "Impact Agenda," government-led research policies in many countries aim to increase the impact of research on society (Gunn & Mintrom, 2016). In various nations, policy interventions promoting the Impact Agenda have fostered a culture in which academics are expected to consider societal benefits. For example, in the UK, university departments are required to describe the non-academic impacts of their research under the UK's Research Excellence Framework (Thelwall, 2021). Such efforts have contributed to the development of altmetrics, which report how often a research paper has been tweeted, shared, or saved on social media, reference management tools, and other online platforms, because traditional citation-based indicators can only reflect impacts within the academic publishing system.

Altmetrics were originally proposed as a form of crowd-sourced peer review to help researchers quickly identify significant scholarly works (Priem et al., 2011). In this sense, altmetric indicators, such as Twitter mentions, are potentially crowd-sourced indicators of research significance. According to Lamont (2009), the concept of significance in research evaluation can be defined in terms of three dimensions: scholarly, political, and social. Therefore, in this study, research significance is understood as the extent to which a study has relevance and impact in one or more of these domains. According to the "wisdom of crowds" concept, in some contexts a diverse group of individuals can make better decisions and predictions than a few individuals or even an expert (Galton, 1907; Hertwig, 2012; Lorenz et al., 2011; Surowiecki, 2005). But not all crowds are "wise." One of the criteria required to form a wise crowd is independence (Surowiecki, 2005). Without independence, much of or at least part of the information selected can occur by crowd following rather than through separate judgments, and in this case the crowd selection process may not lead to better outcomes. Banerjee (1992) referred to sequential decision making behaviour involving crowd following as "herd behaviour," which could lead to a cascading failure, called the "stupidity of herds" by Yu et al. (2018).

Herd behaviour is common in many social and economic situations (Banerjee, 1992). In the two-step flow of communication model, Katz (1957) proposed that, in general, ideas first flow from mass media to opinion leaders, who in turn spread them to a segment of the population that they influence. Even on social media, where ideas seem to flow directly between peers, the influence of opinion leaders has been found in several areas. Numerous studies consider retweets on social media as indicators of social influence (Cano et al., 2014; Yin & Zhang, 2012; Zhang et al., 2015). If retweets signify social influence, Twitter data greatly violates the independence assumption above because there are more retweets than original tweets (Firdaus et al., 2018; Shioda & Minamikawa, 2019). In addition, the distribution of retweets follows a power law (Lu et al., 2014), meaning that not all the original tweets or tweeters are equally retweeted. In particular, a few influential tweeters generate most retweets, and content popularity on Twitter can therefore be due to these influencers.

Power law and similar distributions can be created by cumulative advantage mechanisms, also known as the Matthew effect (Perc, 2014). This has been hypothesised to occur for academic citations because people are more likely to cite highly cited papers because they are better known, or for other reasons not related to the quality of the paper. The Matthew effect in citation counts has led to criticism of the use of citation counts as indicators of research significance, as they unfairly reward researchers with greater social capital or fame rather than reflecting the actual quality or impact of their work. However, Wang (2014) argued that while there are components in the Matthew effect that challenge the validity of citation counts as indicators of research importance, the usefulness of a paper to others, such as to provide foundations, does not undermine the validity of citation counts as valid indicators of research importance. While the Matthew effect and the underlying influences, and therefore the validity of citation counts as indicators of research significance, are well understood in citation counts, they remain largely unexplored in altmetric research.

This study contributes to this area by investigating whether social influence and the Matthew effect, manifested in retweets, in Twitter mentions of research papers challenges the validity of Twitter mentions (tweet counts and tweeter counts) as crowd-sourced altmetric indicators of research significance.

2. LITERATURE REVIEW

2.1. Wisdom of Crowds or Stupidity of Herds

Social media offers unprecedented opportunities for the development of crowd-sourced recommendation solutions. Research on the wisdom of crowd effect indicates that when certain conditions are met, the collective judgments of a group can be more accurate than the judg-

 https://doi.org/10.1633/JISTaP.2025.13.1.1

ments of individuals (Galton, 1907; Hertwig, 2012; Lorenz et al., 2011; Surowiecki, 2005). Azar and Lo (2016) showed the wisdom of Twitter crowds by using tweet data to successfully predict stock market returns. One crucial requirement for the wisdom of crowd effect is independence (Surowiecki, 2005). A study by Lorenz et al. (2011) found that social influence can reduce the wisdom of crowd effect over time. This occurs because social influence tends to cause individuals' estimates to converge, decreasing group opinion diversity. Consequently, the group's overall accuracy decreases due to the loss of independent judgments.

Golub and Jackson (2010) argued that a society is "wise," meaning all opinions in a society converge to the truth, only when the influence of the most influential agent vanishes as the society grows. For example, if individuals listen to only one particular agent, their beliefs will then converge to that agent's initial information and have a substantial probability of deviating substantially from the truth. When individuals follow the actions or decisions of others, regardless of their private information, it is known as herd behavior, as identified by Banerjee (1992). Yu and Chen (2021) explored the impact of big data on the quality of society-level decision-making. The study revealed that when big data technology reduces the attention cost only for shallow information, but not for deep information, it leads to an increase in herd behaviour among individuals. This results in inadequate exploitation of available information and ultimately leads to a decline in the quality of society-level decisions.

Social influence on social media platforms has been extensively studied, with retweets often used as indicators of influence (Venkatesan et al., 2021; Yin & Zhang, 2012). According to Venkatesan et al. (2021), a retweet reflects the influence of the original tweet as well as the original tweet author, as the author was cited along with the content. Prior research on the dynamics of tweeters suggests that retweeting is the core driver of content popularity on Twitter, and there are more retweets than original tweets about real-world phenomena (Firdaus et al., 2018; Shioda & Minamikawa, 2019). Shioda and Minamikawa (2019) also found that retweets accounted for 65% to 86% of all tweets discussing real-world news events, such as "earthquake" and "Miss Universe." For health-related topics, Albalawi et al. (2019) collected two datasets of tweets containing health-related Arabic keywords, each of which included around 200,000 tweets. In their datasets, original tweets comprised 20% to 27% of all tweets.

A similar trend was found in an early dataset of COVID-19 tweets. Out of 74,758 tweets, only about 12% were original tweets and the remaining 88% were retweets (Prager et al., 2021). A frequency analysis of tweets relating to COVID-19 "intubation box(es)," "aerosol box(es)," or "intubation barrier(s)" indicated that, for 3,899 original tweets, there were 18,831 retweets (Mariano et al., 2021). But in a rare case where original tweet counts were slightly higher than retweet counts, Yoon et al. (2019) reported that 52% of all tweets were original in their study of the Zika virus. For altmetrics, Pandian et al. (2019) studied Twitter mentions of psychology research articles and found that retweeting increases the number of Twitter mentions received by articles. Although previous studies have suggested that the popularity of content on Twitter depends on retweets and most tweets are retweets, the proportions of original tweets/tweeters and retweets/retweeters in the overall tweet or tweeter counts of research papers are largely unexplored in general, and in the medical subject area in particular. The following research question addresses this gap:

RQ1: To what extent is the tweeting of COVID-19 medical research papers an independent process in terms of the proportions of original tweets/tweeters and retweets/retweeters?

2.2. The Matthew Effect in Citation Counts and Twitter Mentions

In addition to the findings that most tweets are retweets, the distribution of retweet counts tends to be highly skewed, following a power law (Mathews et al., 2017): Not all original tweets or tweeters are equally retweeted, but retweeting is concentrated on a few. In altmetrics research, Pandian et al. (2019) found that the popularity of psychology research articles on Twitter was primarily determined by their most popular tweet. To understand the challenge imposed by the power-law distribution of retweet counts on the validity of Twitter mentions as indicators of research significance, we first need to understand the similar phenomena in citation counts. Merton (1973, p. 446) defined the Matthew effect in science as "the accruing of greater increments of recognition for particular scientific contributions to scientists of considerable repute and the withholding of such recognition from scientists who have not yet made their mark."

Wang (2014) identified three components in the Matthew effect in citations: Networking, prestige, and appropriateness. For the networking effect, Blashfield (1982, p. 3) found that a highly cited paper was mostly cited by

closely connected authors, known as the "invisible college," which is "a collection of intellectuals who had a sense of allegiance to each other and who frequently interacted both professionally and socially." Several studies have also provided strong evidence of the influence of the author, institution, and journal prestige on citation counts (Larivière & Gingras, 2010; Medoff, 2006; Tanner-Smith & Polanin, 2016). Due to the networking and prestige components of the Matthew effect, research evaluation based on citations has been criticized as rewarding fame or social capital rather than the intrinsic value of a paper, and so "citation-based metrics should not be used as fair tools for research evaluations" (Wang, 2014, p. 329). Appropriateness, in the sense of the usefulness of a paper to others, such as providing foundations that others can build their career on, is a third component of the Matthew effect (Wang, 2014). The appropriateness effect may also be cumulative, as the more people pursue the problems raised by the paper, the more opportunities for further enquiries and the higher the appropriateness of the paper. When discussing the Matthew effect in citations, it is the networking and prestige components that challenge the validity of citation metrics, not the appropriateness component (Wang, 2014).

While the distribution of altmetric data is known to follow power law (Banshal et al., 2022), the distribution of retweets in altmetric data is largely unexplored in many areas in altmetrics research, including for medical research. This leads to the next research question:

RQ2: Does the distribution of retweets suggest the existence of skewing in the data, in that some original tweeters are retweeted disproportionately more than others?

2.3. Credibility Heuristic Cues of Original Tweeters and Retweeting

One way to find conditions that lead to a power law or similar highly skewed distributions of retweet counts is to identify the characteristics of original tweeters (i.e., the authors of the original tweets) that are associated with retweet rates. One Twitter user characteristic known to be linked to retweeting and information sharing on Twitter is source credibility (Boehmer & Tandoc, 2015; Son et al., 2020). In the traditional sense, the source of a research article is the journal in which it is published. But on social media and the Internet, multiple layers of sources are usually involved in the spread of information. On Twitter, in addition to the credibility of the tweeter and the tweet content, studies based on dual process theories suggest that some peripheral cues (e.g., academic credentials) of a visible source (e.g. the user tweeting the content) also influence audience perceptions and decisions to retweet (Son et al., 2020).

Dual process theories, such as the heuristic-systematic model (HSM), suggest that people process information in a dual fashion, in systematic and heuristic modes (Chaiken, 1980). In the systematic mode, people expend considerable cognitive effort to analyse content to form judgments. In contrast, while in heuristic mode, they rely on easily available peripheral "cues" about the content to form quick judgements. Here, "heuristics" are generalizations of knowledge developed from previous experience (Chen & Chaiken, 1999), and "cues" serve to trigger them. For example, "experts' statements can be trusted" is an expertise heuristic that can be invoked by the expert source cue in the presentation of the message (Sundar, 2008).

Taking the heuristic approach, the MAIN model of Sundar (2008) argues that technology can assimilate information about multiple sources in the communication process that are influential in shaping users' perceptions and credibility assessments of sources and the underlying content. In the MAIN model, Sundar (2008) proposed four broad technology affordances: Modality, Agency, Interactivity, and Navigability, each of which is a repository of cues that have been demonstrated to trigger cognitive heuristics for credibility judgements. Agency affordances of technology are cues attributed to sources involved in the chain of online communication that trigger corresponding heuristics, which in turn affect audience perceptions about the credibility of the information provided by them. For example, users make quick judgments about the credibility of a social media post or an online website via the expertise heuristic, cued when they see an expert as the source of information.

On Twitter, account authoring tweets were investigated as sources of information, and the authority, identity, and bandwagon cues of those sources were found to be linked to credibility perceptions of the audience, supporting the MAIN model (Lin et al., 2016). Son et al. (2020) empirically examined the influence of reputation, social presence, and recency cues of sources on retweet decisions, finding strong evidence for the importance of tweeter profile information in retweet decisions during disastrous situations. More interestingly, Remy et al. (2013) found that the parameters of the power-law distribution of retweets can be estimated by the follower count of Twitter users, which is a bandwagon cue for tweeters that invokes bandwagon heuristics associated with credibility perception,

 https://doi.org/10.1633/JISTaP.2025.13.1.1

according to the MAIN model. Their study suggested that the more followers a user has, the higher the probability of their tweets being highly retweeted, triggering power-law behaviour in retweets.

Thus, to better understand how original tweeters influence the retweeting of research papers, which ultimately affects the overall tweet count or tweeter counts discussing the papers, this research investigates the relationship between the credibility heuristic cues of the original tweeters and their retweet counts. Specifically, this study examined the following cues of original tweeters in relation to the retweet counts they receive.

2.3.1. Authority Cues

The influence of source authority cues on audience credibility perceptions has been extensively demonstrated in the literature. In one study, Lin et al. (2016) found that authority cues affected audience source credibility perceptions most strongly in the case of tweets containing health risk information related to the rise of drug-resistant gonorrhoea. As for COVID-19 information sharing intentions, Lu et al. (2021) found that source trust was also positively associated with information sharing intention, and health professionals were the most-trusted source, followed by academic institutions, government agencies, and news media. This leads to the following hypothesis:

H1: Authority cues of original tweeters positively associate with their retweet counts.

2.3.2. Bandwagon Cues

This cue is based on the psychology of "other users" and collective endorsement. Sundar (2008) claimed that people tend to have a favourable attitude toward products or content endorsed by many people. Thus, bandwagon cues are mainly popularity indicators, such as the number of followers, through which credibility is established. Lee and Sundar (2013) tested the number of followers that a source has as a bandwagon cue associated with credibility assessments, and found that it was associated with the perceived credibility of health content. In a separate study on a different topic, more-followed tweeters were found to be more likely to be retweeted (Son et al., 2020). This leads to the following hypothesis:

H2: Bandwagon cues of original tweeters positively associate with their retweet counts.

2.3.3. Social Presence Cues

In times of crisis, such as the global outbreak of a disease, social media serves not only as an information-sharing tool but also as a place to get moral support (Mukkamala & Beck, 2017). Consistent posting is known to be favoured by social media algorithms (Teah Hopper Consulting, 2018), and the more people see someone's tweets, the stronger the sense of togetherness everyone feels (Son et al., 2020). For example, Jahng and Littau (2016) found that highly interactive journalists were perceived as more credible than their less interactive counterparts on Twitter. Son et al. (2020) found significant relationships between social presence cues (status count and affiliation length) and retweet rates on Twitter. This leads to the following hypothesis:

H3: Social presence cues of original tweeters positively associate with retweet counts.

In addition, previous studies have indicated that there is an interaction effect between bandwagon cues and authority cues on the credibility of messages (Kim & Sundar, 2016; Lee & Sundar, 2013). Lee and Sundar (2013) studied the perceived credibility of health messages on Twitter and found that the presence of authority cues (professional accounts) with large numbers of followers (bandwagon cues) was more credible. In contrast, Kamiński et al. (2021) found that when it comes to tweets on COVID-19, those from celebrities with the largest number of followers were retweeted the most, outperforming those from health and scientific organizations, indicating that authority cues alone may not lead to high retweet rates. Therefore, interaction effects between authority cues (present vs. absent) and bandwagon cues (follower counts) on retweet counts were also tested in the present study. It is expected that among the original tweeters with authority cues, those with high bandwagon cues (high follower counts) will be retweeted more.

H4: Bandwagon cues moderate the relationship between authority cues and retweet counts.

3. METHODS

COVID-19 articles from a single major medical journal, *The Lancet*, were selected to control for journal and topic-related influences on the number of tweets and retweets received by the articles. Due to the importance of rapidly disseminating medical research about COVID-19

on Twitter, this topic should generate high Twitter mentions, giving higher statistical power to the study, although it will not generalise. Previous literature has indicated that the availability of data, and thus the potential usefulness of the altmetrics, varies across subject areas and journals (Htoo & Na, 2017; Thelwall et al., 2013). More Twitter interactions occur for articles from medical subject areas (Thelwall et al., 2013) and for articles from higher ranking journals (Htoo & Na, 2017), supporting the likely high volume of data. *The Lancet* is a well-known and important general medical journal. It was the highest-ranking journal in the medicine (general & internal) category, according to the 2021 Journal Impact Factor and so was selected in preference to other similarly important general medical journals.

3.1. Data Collection

The Scopus keyword query "COVID-19," matching the title, abstract or keyword fields of articles from 2020 with the results filtered to *The Lancet,* identified 53 articles (document type=article). All tweets (tweet IDs) mentioning these 53 articles were collected from Altmetric.com via its applications programming interface (API) by matching article DOIs. There were 246,840 tweet IDs collected from Altmetric.com for the 53 articles. For each tweet ID, the corresponding tweet and user information was retrieved from Twitter via its API. This included the retweet count, retweeted status (to differentiate between original tweets and retweets), and for retweets, the original tweet ID and original tweeter details. For tweet IDs that were no longer valid, no information was collected. Details for 238,339 tweets (177,011 retweets and 61,328 original tweets) for 53 articles were successfully collected from Twitter.

Unfortunately, Altmetric.com's coverage was found to be incomplete in the data collected via its API. For instance, one original tweet (tweet ID: 1285207186591887360) had 9,863 retweets according to Twitter, but only 2,050 were associated with it in the data collected from Altmetric.com's API, revealing a significant gap in its coverage. Additionally, some tweets included in the data from Altmetric.com were no longer available on Twitter, further limiting its coverage. These limitations are not unique to this study, but apply to all studies using data from Altmetric.com and likely other altmetric data providers as well.

3.2. Sampling Original Tweeters

To address RQ3 and hypotheses H1, H2, and H3 concerning the nature of the influence of original tweeters, a sample of original tweeters was selected for a manual content analysis to help find the relationship between their retweet counts and their credibility cues.

Among 177,011 retweets in the dataset, 867 (0.49%) were self-retweets. Although this is a very small number and unlikely to affect the results, self-retweets were filtered out before sampling original tweeters. After filtering out self-retweets, there were 176,145 retweets in the dataset. For each retweet in the dataset, Twitter provides a corresponding original tweet ID, retweet count, and original tweeter information. There were 332 different retweet counts, giving 332 groups of original tweeters. The group with the highest retweet count (9,863) had one member (*The Lancet*). In the groups with the lowest retweet counts (1 and 0), there were 4,494 members with 1 retweet and 4 members with no retweets. When sampling original tweeters, one tweeter from each group was selected. In groups with more than one user, one user was selected randomly. The same user was selected only once. For users appearing in more than one group, their average retweet count was calculated and used for analysis. This process produced 265 original tweeters for analysis.

3.3. Measures of Credibility Heuristics Cues

3.3.1. Authority Cues

Perceived credibility is often investigated with heuristic cues based on official, professional, and academic statuses. Lee and Sundar (2013) explored the effects of three authority cues on the perceived credibility of tweeted health messages. They used the affiliation "doctor" (professional status) as a high-authority cue. Similarly, Zhang et al. (2021) explored how users judge information quality in academic social networking sites and used a doctorate or master's degree (academic status) as a high authority cue. Likewise, Lin et al. (2016) evaluated the effects of three agency cues (authority, identity, and bandwagon) on the perceived credibility of tweets about drug-resistant gonorrhoea. They operationalized The Center for Disease Control and Prevention (CDC) as the expert authority cue, given that it is an "official" Twitter account (official status) associated with the topic.

When it comes to authoritative information, journalists and the media are also well established, although their status may have eroded recently. When examining the role of source credibility in the vaccination debate, Stueckemann (2019) found that well-established news outlets had higher credibility evaluations. Based on the previous literature, this study uses health officials, health professionals, academic, and established media sources as authority cues,

 https://doi.org/10.1633/JISTaP.2025.13.1.1

as described in the Appendix. Original tweeters in the sample were coded by two independent coders using the coding instructions in the Appendix. Interrater reliability, estimated using Cohen's kappa, was 0.7 for the medical category, 0.7 for the academic category, 0.4 for the government/public health/medical official category, and 0.7 for the news media category. Table 1 shows the frequency of authority cues by category.

The kappa results were interpreted as: Values ≤0 no agreement, 0.01-0.20 none to slight, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and 0.81-1.00 almost perfect agreement (Landis & Koch, 1977). Cohen's suggested interpretation implies that a score as low as 0.41 might be acceptable (McHugh, 2012). The kappa values for all categories, except for government/public health/medical official, are above 0.41. Since the government/public health/medical official category made up only 5%, it is unlikely that it will affect overall results. Therefore, all the categories are included in the analysis.

3.3.2. Bandwagon Cues

Adopting the measures by Lee and Sundar (2013), the number of followers is used in this study as a bandwagon cue.

3.3.3. Social Presence Cues

Adopting the measures by Son et al. (2020), status count (number of posts) and affiliation length/age of account (number of days since the account was created up to February 11, 2021) are used in this study as social presence cues. We chose February 11, 2021, as the cut-off date since it was the most recent date available in the dataset. Hence, all accounts in the dataset have an affiliation length or age of 0 or greater than 0.

3.4. Analysis

As supported by prior studies, retweets were considered an indicator of social influence in this study. Thus, RQ1 was addressed by finding the proportion of original tweets/tweeters and retweets/retweeters, and for RQ2 by plotting the distribution of retweet count to original tweets/tweeters in the dataset using the poweRlaw R package (Gillespie, 2015). Since a minimum retweet value of 1 is required to plot a power-law distribution, 1 was added to all retweet counts when plotting the retweet distribution. To investigate the effect of credibility heuristics cues on retweet counts, the Mann-Whitney U test and multiple linear regression were selected. Because the dependent variable retweet count was highly skewed, the Mann-Whitney U test was used to test the difference in median retweet counts between the two groups: The one with authority cues and the one without.

When choosing a suitable regression model in citation and altmetrics studies, as the data is skewed and the variance is greater than the mean, some altmetrics studies have used negative binomial multiple regression (Pandian et al., 2019). However, Thelwall and Wilson (2014) showed that ordinary least squares regression, after a log transformation, is the most suitable regression strategy for citation and altmetric data as it takes into account very high values, which are typical for skewed distributions. Thus in this study, multiple linear regression was conducted with log-transformed retweet counts ln(1+retweet count) as the dependent variable. Continuous independent variables (follower count, number of tweets posted, and number of days since the account was created) were mean-centred to make the results more interpretable, and to avoid multicollinearity when testing for interaction effects.

Table 1. Frequencies of authority cues by category (n=265)

Category	n (%)
Medical	77 (29)
Academic	115 (43)
Public health	13 (5)
Media	41 (16)

4. RESULTS AND DISCUSSION

4.1. Wisdom of Crowds or Stupidity of Herds

Out of 238,339 tweets for 53 articles, 61,328 (26%) were original tweets tweeted by 38,172 unique original tweeters, and the remaining 177,011 (74%) were retweets tweeted by 125,264 unique retweeters. All 53 articles had received at least 22 tweets, perhaps from publicity by the journal's Twitter account @TheLancet, which had around 724,000 Twitter followers as of August 2022. There were varying proportions of retweets between articles, from 40% to 93% (Fig. 1), in line with the general pattern of tweets and retweets ratios discussed above. Addressing RQ1, the results confirmed that tweeting research papers, at least for medical research papers relating to COVID-19, is not an independent process. In fact, the extent of influence, as revealed by the proportion of original tweets/tweeters and retweets/retweeters in total tweets/tweeters, is substantial. Fortunately, it is easy to address this issue in

tweet/tweeter counts as altmetric indicators. In addition to reporting the total number of tweets/tweeters, as done by altmetric data aggregators like Altmetric.com, the number of original tweets/tweeters could also be reported.

4.2. Influence of Original Tweeters

Distributions of retweet counts for both original tweets and original tweeters were generated. Fig. 2A shows the distribution of retweet counts for original tweets, and the line follows a power law reasonably well (exponent=1.89) with p=0.43 for a goodness-of-fit test, failing to reject the null hypothesis that data is generated from a power-law distribution.

Since some original tweeters tweeted the same paper more than once, original tweeters were consolidated for each paper by taking their average retweet count for the same paper. After consolidation, there were 11,892 original tweeters associated with 177,011 retweets retrieved from Altmetric.com. Fig. 2B shows the distribution of retweet counts for 11,892 original tweeters, and the line has a power-law fit (exponent=1.89) with p=0.37 for a goodness-of-fit test, failing to reject the null hypothesis that data is generated from a power-law distribution. The distributions of retweets for both original tweets and original tweeters are very similar since most data points are the same in both cases. Overall, it is clear that a large majority of retweets were from a few highly retweeted influential original tweeters, addressing RQ2.

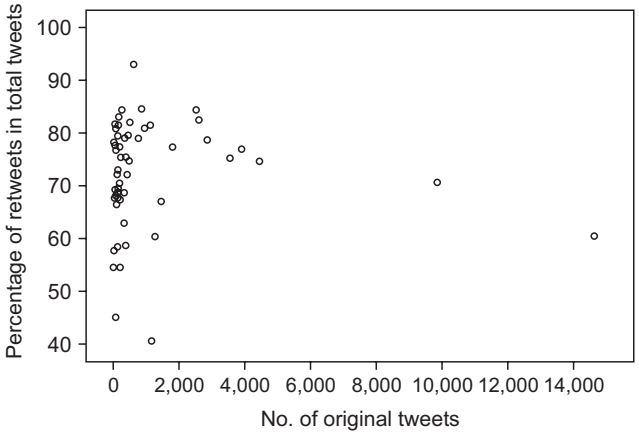


Fig. 1. Proportion of retweets vs. the number of original tweets received by the 53 papers investigated.

4.3. Credibility Heuristic Cues of Original Tweeters and Retweeting

Out of the 265 manually classified original tweeters, 72% had authority cues (belonging to at least one category in the Appendix) and 28% did not. Mann-Whitney U testing was applied to determine if there were differences in retweet counts for original tweeters with and without authority cues. The median retweet counts for the two groups were not significantly different (Table 2), rejecting H1.

Next, hierarchical multiple linear regression was conducted to test the main effects of all independent variables as well as interaction effects between authority cues (present vs. absent) and bandwagon cues (follower count) on logged retweet counts. Descriptive statistics for the variables are shown in Table 3. In the first step, all inde-

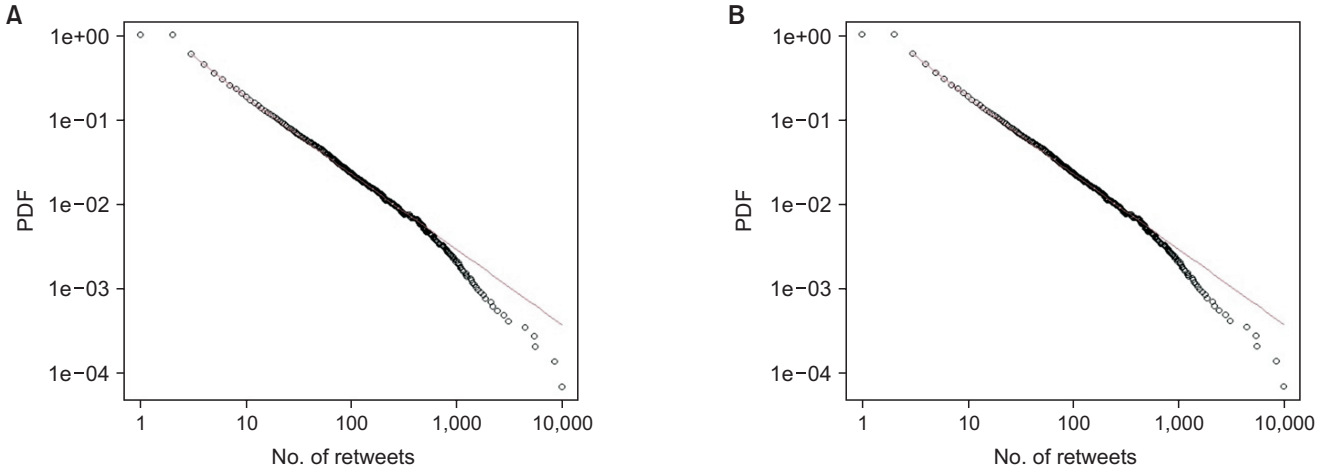


Fig. 2. Distribution of retweets. (A) Distribution of retweet counts per original tweet (probability distribution function [PDF] with a power-law fit [solid line] [exponent=1.89]). (B) Distribution of retweet counts per original tweeter (PDF with a power-law fit [solid line] [exponent=1.89]).

 https://doi.org/10.1633/JISTaP.2025.13.1.1

Table 2. Mann-Whitney U test results for retweet counts of original tweeters with and without authority cues

Group	n (%)	Median retweet count	Mean rank	Z	Mann-Whitney U	*p*
Users with authority cues	192 (72)	82.83	134.75	-0.60	6,672.50	0.55
Others	73 (28)	92	128.40			

Table 3. Descriptive statistics of variables

	Minimum	Maximum	Skewness	Mean	Medium	Standard deviation
Dependent variable						
Retweet count (original variable not included in the analysis)	0	2,826	4.23	198.10	85	330.70
Logged retweet count (transformed variable used in the analysis)	0	7.95	-0.11	4.46	4.45	1.33
Independent variables						
Follower count	5	7,964,976	9.57	179,124.99	36,784	601,546.43
Status count	178	474,770	3.24	44,836.74	19,923	70,127.90
Age of account	0	5,102	-0.75	3,123.76	3,410.76	1,279.04

Table 4. Results of hierarchical regression analysis

Step	Predictors	Standardized coefficients		R^2/R^2 change	*F*	*p*	Tolerance/variance inflation factor
		β	*p*				
1				0.094/0.094	6.70	<0.001	
	Authority	0.03	0.612				0.954/1.04
	Follower count	0.248	<0.001				0.983/1.02
	Status count	-0.152	0.018				0.862/1.16
	Age of account	0.131	0.037				0.889/1.13
2				0.11/0.01	6.181	<0.001	
	Authority	0.024	0.689				0.951/1.051
	Follower count	0.581	0.001				0.105/0.105
	Status count	-0.144	0.024				0.859/1.165
	Age of account	0.121	0.055				0.882/1.134
	(Follower count* Authority)	-0.351	0.053				0.106/9.446

*The interaction between variables.

pendent variables were entered as independent predictors of logged retweet counts (Table 4). In the second step, the interaction effect was tested by entering the product of the authority cues (present vs. absent) and mean-centred follower count as an additional predictor (Table 4).

The regression results showed that both bandwagon cues (follower count) and social presence cues (status count and age of account) are significant predictors of retweet counts at a significant level p=0.05 in step 1 (Table 4), supporting H2 and H3. The strongest predictor of higher retweet rates was the follower count. Although the main effect of authority cues was not significant, the interaction effect between authority cues and bandwagon cues (follower counts) was close to significant with a p-value

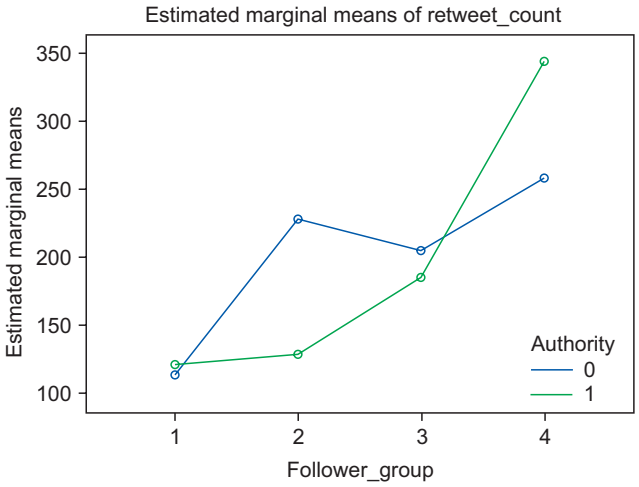


Fig. 3. Interaction between authority cues and bandwagon cues (follower count) on retweet counts.

of 0.053 in step 2 (Table 4), providing marginal support for H4. To plot the interaction effect, the follower count variable was recoded as a categorical variable with 4 levels. The original tweeters in the first quartile with the lowest number of followers were coded as 1, those in the second quartile as 2, the third quartile as 3, and the fourth quartile with the highest number of followers was coded as 4.

As shown in Fig. 3, in the presence of authority cues, the higher the follower count of original tweeters, the more retweets they receive. The effect of authority cues is strongest for original tweeters with the highest follower count and weakest for those with the lowest follower count. In contrast, for original tweeters with no authority cues, the retweet rate does not seem to follow the pattern of follower count consistently. But among them, in general, the top 75% with the highest follower count received similar retweet rates. Although H1 is not supported, based on the findings supporting H2, H3, and H4, it can be concluded that, overall, the results offered evidence for credibility cues of original tweeters as significant factors associated with retweet rates, addressing RQ3. These results support previous research that highlights the role of credibility heuristic cues in shaping the magnitude of retweets received by tweeters.

5. CONCLUSIONS

Although based on a single topic, year, and journal, this study found evidence of social influence, as manifested by retweets, in the overall tweet/tweeter counts of research papers. The high proportion of retweets and the power-law-like distribution of retweets to original tweets and tweeters in the data is consistent with herd-like behaviour and against the wisdom of crowd effect. However, as for citation counts, not every condition giving rise to a power law challenges the validity of the indicators. Further analysis on the characteristics of original tweeters associated with retweets revealed that the retweet rate was associated with the credibility cues of original tweeters, which are associated with the credibility perception of the retweeters. If retweets are mostly positively associated with credible original tweeters with the expertise to identify important works in the field, retweets can be seen as reinforcements of the importance of the papers to retweeters. In this case, even if Twitter mentions are heavily influenced by retweets, this may not undermine their validity as indicators of research significance.

Examining the findings from another perspective, associations between credibility heuristic cues of original tweeters and retweet rates suggest that retweets are more likely influenced by the perceived credibility of the original tweeters rather than the quality of the research papers they share. This raises an important question: Given that a large majority of tweets are retweets, if retweets are influenced by the credibility of the original tweeters rather than the quality/significance of research papers, can tweets be considered valid indicators of the quality/significance of research papers? They are, evidently, more directly indicative of the extent to which credible tweeters have tweeted the papers. The findings of this research support separating original tweets from retweets, as previously suggested by Fang et al. (2020) and Haustein (2019). The current research goes further by revealing the need to re-evaluate the use of retweets as indicators for research papers. The implicit assumption for including retweets within altmetric indicators is that retweets reflect the significance of research papers to some extent. The current paper's suggestion that they may more directly reflect the credibility of the original tweeters undermines this claim to some extent.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Albalawi, Y., Nikolov, N. S., & Buckley, J. (2019). Trustworthy health-related tweets on social media in Saudi Arabia:

Tweet metadata analysis. *Journal of Medical Internet Research,* 21(10), e14731. https://doi.org/10.2196/14731

Azar, P. D., & Lo, A. W. (2016). The wisdom of Twitter crowds: Predicting stock market reactions to FOMC meetings via Twitter feeds. *Journal of Portfolio Management,* 42(5), 123-134. https://doi.org/10.3905/jpm.2016.42.5.123

Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics,* 107(3), 797-817. https://doi.org/10.2307/2118364

Banshal, S. K., Gupta, S., Lathabai, H. H., & Singh, V. K. (2022). Power laws in altmetrics: An empirical analysis. *Journal of Informetrics,* 16(3), 101309. https://doi.org/10.1016/j.joi.2022.101309

Blashfield, R. K. (1982). Feighner et al., invisible colleges, and the Matthew effect. *Schizophrenia Bulletin,* 8(1), 1-6. https://doi.org/10.1093/schbul/8.1.1

Boehmer, J., & Tandoc, E. C., Jr. (2015). Why we retweet: Factors influencing intentions to share sport news on Twitter. *International Journal of Sport Communication,* 8(2), 212-232. https://doi.org/10.1123/ijsc.2015-0011

Cano, A. E., Mazumdar, S., & Ciravegna, F. (2014). Social influence analysis in microblogging platforms – A topic-sensitive based approach. *Semantic Web,* 5(5), 357-372. http://doi.org/10.3233/SW-130108

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology,* 39(5), 752-766. https://doi.org/10.1037/0022-3514.39.5.752

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken, & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73-96). Guilford Press.

Fang, Z., Dudek, J., & Costas, R. (2020). The stability of Twitter metrics: A study on unavailable Twitter mentions of scientific publications. *Journal of the Association for Information Science and Technology*, 71(12), 1455-1469. https://doi.org/10.1002/asi.24344

Firdaus, S. N., Ding, C., & Sadeghian, A. (2018). Retweet: A popular information diffusion mechanism – A survey paper. *Online Social Networks and Media,* 6, 26-40. https://doi.org/10.1016/j.osnem.2018.04.001

Galton, F. (1907). Vox populi. *Nature,* 75(1949), 450-451. https://doi.org/10.1038/075450a0

Gillespie, C. S. (2015). Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software,* 64(2), 1-16. https://doi.org/10.18637/jss.v064.i02

Golub, B., & Jackson, M. O. (2010). Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics,* 2(1), 112-149. https://doi.org/10.1257/mic.2.1.112

Gunn, A., & Mintrom, M. (2016). Higher education policy change in Europe: Academic research funding and the impact agenda. *European Education,* 48(4), 241-257. https://doi.org/10.1080/10564934.2016.1237703

Haustein, S. (2019). Scholarly Twitter metrics. In W. Glänzel, H.F. Moed, U.Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 729-760). Springer. https://doi.org/10.1007/978-3-030-02511-3_28

Hertwig, R. (2012). Tapping into the wisdom of the crowd—With confidence. *Science (New York, N.Y.),* 336(6079), 303-304. https://doi.org/10.1126/science.1221403

Htoo, T.H.H. & Na, J. C. (2017). Disciplinary differences in altmetrics for social sciences. *Online Information Review*, 41(2), 235-251. https://doi.org/10.1108/OIR-12-2015-0386

Jahng, M. R., & Littau, J. (2016). Interacting is believing: Interactivity, social cue, and perceptions of journalistic credibility on Twitter. *Journalism & Mass Communication Quarterly,* 93(1), 38-58. https://doi.org/10.1177/1077699015606680

Kamiński, M., Szymańska, C., & Nowak, J. K. (2021). Whose Tweets on COVID-19 gain the most attention: Celebrities, political, or scientific authorities? *Cyberpsychology, Behavior and Social Networking,* 24(2), 123-128. https://doi.org/10.1089/cyber.2020.0336

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly,* 21(1), 61-78. https://doi.org/10.1086/266687

Kim, H. S., & Sundar, S. S. (2016). Motivating contributions to online forums: Can locus of control moderate the effects of interface cues? *Health Communication,* 31(5), 583-595. https://doi.org/10.1080/10410236.2014.981663

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Harvard University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 33(1), 159-174. http://doi.org/10.2307/2529310

Larivière, V., & Gingras, Y. (2010). The impact factor's Matthew effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology,* 61(2), 424-427. https://doi.org/10.1002/asi.21232

Lee, J. Y., & Sundar, S. S. (2013). To tweet or to retweet? That is the question for health professionals on Twitter. *Health Communication,* 28(5), 509-524. https://doi.org/10.1080/10410236.2012.700391

Lin, X., Spence, P. R., & Lachlan, K. A. (2016). Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior,* 63, 264-271. https://doi.org/10.1016/j.chb.2016.05.002

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America,* 108(22), 9020-9025. https://doi.org/10.1073/pnas.1008636108

Lu, L., Liu, J., Yuan, Y. C., Burns, K. S., Lu, E., & Li, D. (2021). Source trust and COVID-19 information sharing: The mediating roles of emotions and beliefs about sharing. *Health Education & Behavior,* 48(2), 132-139. https://doi.org/10.1177/1090198120984760

Lu, Y., Zhang, P., Cao, Y., Hu, Y., & Guo, L. (2014). On the frequency distribution of retweets. *Procedia Computer Science,* 31, 747-753. https://doi.org/10.1016/j.procs.2014.05.323

Mariano, E. R., Kou, A., Stiegler, M. A., & Matava, C. (2021). The rise and fall of the COVID-19 aerosol box through the lens of Twitter. *Journal of Clinical Anesthesia,* 69, 110145. https://doi.org/10.1016/j.jclinane.2020.110145

Mathews, P., Mitchell, L., Nguyen, G. T. K., & Bean, N. G. (2017, April 3-7). The nature and origin of heavy tails in retweet activity. *Proceedings of the 26th International World Wide Web Conference* (pp. 1493-1498). International World Wide Web Conferences Steering Committee.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica,* 22(3), 276-282.

Medoff, M. H. (2006). Evidence of a Harvard and Chicago Matthew effect. *Journal of Economic Methodology,* 13(4), 485-506. https://doi.org/10.1080/13501780601049079

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

Mukkamala, A., & Beck, R. (2017). *Presence of social presence during disasters*. Paper presented at the 21th Pacific Asia Conference on Information Systems, Langkawi, Malaysia.

Pandian, N. D. S., Na, J. C., Veeramachaneni, B., & Boothaladinni, R. V. (2019). Altmetrics: Factor analysis for assessing the popularity of research articles on Twitter. *Journal of Information Science Theory and Practice,* 7(4), 33-44. https://doi.org/10.1633/JISTaP.2019.7.4.3

Perc, M. (2014). The Matthew effect in empirical data. *Journal of the Royal Society Interface,* 11(98), 20140378. http://doi.org/10.1098/rsif.2014.0378

Prager, R., Pratte, M. T., Unni, R. R., Bala, S., Ng Fat Hing, N., Wu, K., McGrath, T. A., Thomas, A., Thoma, B., & Kyeremanteng, K. (2021). Content analysis and characterization of medical Tweets during the early COVID-19 pandemic. *Cureus,* 13(2), e13594. https://doi.org/10.7759/cureus.13594

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). *Altmetrics: A manifesto*. https://digitalcommons.unl.edu/scholcom/185/

Remy, C., Pervin, N., Toriumi, F., & Takeda, H. (2013, December 2-5). Information diffusion on Twitter: Everyone has its chance, but all chances are not equal. In K. Yetongnon, A. Dipanda, & R. Chbeir (Eds.), *Proceedings of the 2013 International Conference on Signal-Image Technology & Internet-Based Systems* (pp. 483-490). IEEE.

Shioda, S., & Minamikawa, M. (2019, October 22-25). Features found in Twitter data and examination of retweeting behavior. In M. Alsmirat, & Y. Jararweh (Eds.), *Proceedings of the 6th International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 529-534). IEEE.

Son, J., Lee, J., Oh, O., Lee, H. K., & Woo, J. (2020). Using a heuristic-systematic model to assess the Twitter user profile's impact on disaster tweet credibility. *International Journal of Information Management,* 54, 102176. https://doi.org/10.1016/j.ijinfomgt.2020.102176

Stueckemann, E. (2019). *Examining the role of source credibility in the vaccination debate: An experimental study of the influence of heuristic cues on source credibility assessments and attitude change* (master's thesis). Jönköping University, Jönköping, Sweden.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger, & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73-100). MIT Press.

Surowiecki, J. (2005). *The wisdom of crowds.* Anchor Books.

Tanner-Smith, E. E., & Polanin, J. R. (2016). Brief alcohol intervention trials conducted by higher prestige authors and published in higher impact factor journals are cited more frequently. *Journal of Clinical Epidemiology,* 75, 119-125. https://doi.org/10.1016/j.jclinepi.2016.01.028

Teah Hopper Consulting. (2018). *Why posting consistently matters on social media*. https://www.teahhopper.com/blog/2018/11/15/why-posting-consistently-matters-on-social-media

Thelwall, M. (2021). Measuring societal impacts of research with altmetrics? Common problems and mistakes. *Journal of Economic Surveys,* 35(5), 1302-1314. https://doi.org/10.1111/joes.12381

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS One*, 8(5),64841. https://doi.org/10.1371/journal.pone.0064841

Thelwall, M., & Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *Journal of Informetrics,* 8(4), 963-971. https://doi.org/10.1016/j.joi.2014.09.011

Venkatesan, S., Valecha, R., Yaraghi, N., Oh, O. O., & Rao, H. R. (2021). Influence in social media: An investigation of Tweets spanning the 2011 Egyptian revolution. *Manage-*

ment Information Systems Quarterly, 45(4), 1679-1714. https://aisel.aisnet.org/misq/vol45/iss4/6/

Wang, J. (2014). Unpacking the Matthew effect in citations. *Journal of Informetrics,* 8(2), 329-339. https://doi.org/10.1016/j.joi.2014.01.006

Yin, Z., & Zhang, Y. (2012, September 3-5). Measuring pairwise social influence in microblog. In R. Bilof (Ed.), *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (pp. 502-507). IEEE.

Yoon, J., Hagen, L., Andrews, J., Scharf, R., Keller, T., & Chung, E. K. (2019). On the use of multimedia in twitter health communication: Analysis of tweets regarding the zika virus. *Information Research,* 24(2), Article 823. http://informationr.net/ir/24-2/paper823.html

Yu, T., & Chen, S. H. (2021). Big data, scarce attention and decision-making quality. *Computational Economics,* 57(3), 827-856. https://doi.org/10.1007/s10614-018-9798-5

Yu, T., Chen, S. H., & Wang, C. H. (2018). Information aggregation in big data: Wisdom of crowds or stupidity of herds. In E. Bucciarelli, S. H. Chen, & J. Corchado (Eds.), *Decision economics: In the tradition of Herbert A. Simon's heritage* (pp. 16-27). Springer.

Zhang, J., Tang, J., Li, J., Liu, Y., & Xing, C. (2015). Who influenced you? Predicting retweet via social influence locality. *ACM Transactions on Knowledge Discovery from Data,* 9(3), 1-26. https://doi.org/10.1145/2700398

Zhang, N., Yuan, Q., Xiang, X., & Chen, K. (2021). What can you perceive? Understanding user's information quality judgment on academic social networking sites. *Information Research,* 26(2), paper 896. https://doi.org/10.47989/irpaper896

Appendix: Coding Instrument

No.	Categories	Code
1	Medical - Medical professionals & organizations	1 if the account contains cues (words) indicating that the account represents a medical practitioner (such as MD, doctor, Dr.); a medical specialist; or an infectious disease specialist (OR) a medical department/group in a university, a research organization, or a hospital 0 otherwise.
2	Academic -Research/academic professionals & organizations (including academic publishers/ journals)	1 if the account contains cues indicating that the account represents a scientist, researcher, or academic professional or organization (e.g. PhD, scientist, researcher, or lecturer, but not teacher or name of a university). 0 otherwise.
3	Public health - Government officials/Public health professionals & organizations	1 if the account contains cues indicating that the account represents a monarch, president, prime minister, chancellor, minister of health, or head of health departments or organizations (OR) a public health official, professional, or a person with a public health doctorate degree (e.g., Doctor in public health) (OR) A public health-related organization 0 otherwise.
4	Media - Journalists/News organization	1 if the account contains cues indicating that the account represents a professional journalist or another similar professional science communicator (not blogger) (OR) A professional news media organization 0 otherwise.

 https://doi.org/10.1633/JISTaP.2025.13.1.1

JISTaP https://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 13(1): 15-35, 2025

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2025.13.1.2

Topic Modeling Approach on Twitter Data Relevant to Political Changes During the COVID-19 Pandemic

M. G. D. S. Hansika
Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
E-mail: sajinihgamage@gmail.com

K. S. Ranasinghe
Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
E-mail: krsranasinghe@gmail.com

R. A. H. M. Rupasingha*
Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka
E-mail: hmrupasingha@gmail.com

ABSTRACT

The COVID-19 pandemic has affected various sectors of society, including politics. The political changes have had both positive and negative impacts on people's lives. Different public discussions happened during that situation on social media. It is essential to understand those discussions to prepare for the same kind of situation in future. Therefore, this study aims to identify the topics discussed on Twitter regarding this influence. During March 2020 and December 2021, 10,658 Tweets were gathered through the Twitter application programming interface and preprocessed using Python libraries. After feature extraction using the bag-of-words method, both probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) were used as topic modeling methods. As a result of the analysis, 15 topics by LDA and 25 topics by PLSA were extracted during the study and then grouped into five key themes: Government responses for managing the COVID-19 Pandemic, Government decisions for COVID-19, Public response to government measures for COVID-19, Social influence, and Vaccination. Through a comparative evaluation of the LDA and PLSA topic modeling techniques, the research identifies LDA as the superior method, providing more accurate and coherent results.

Keywords: COVID-19, political changes, topic modeling, Twitter, public opinion

Received: January 18, 2024 **Revised:** September 19, 2024
Accepted: October 15, 2024 **Published:** March 30, 2025

***Corresponding Author:** R. A. H. M. Rupasingha
https://orcid.org/0000-0003-3922-4290
E-mail: hmrupasingha@gmail.com

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. Open Access articles are automatically archived in the Korea Institute of Science and Technology Information (KISTI)'s Open Access repository (AccessON). All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/). Under this license, the authors retain full ownership of their work, while permitting anyone to use, distribute, and reproduce the content in any medium, as long as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© 2025 [M. G. D. S. Hansika, K. S. Ranasinghe, R. A. H. M. Rupasingha] This article is distributed under the Creative Commons Attribution License (CC BY), allowing unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. It is published by the Korea Institute of Science and Technology Information (KISTI).

1. INTRODUCTION

Epidemics are not new to humans. Humanity has already been affected by major epidemics such as Ebola, the bird flu virus, and the H1N1 influenza pandemic (Lim, 2021). However, the coronavirus disease (COVID-19) has caused more impact than any other virus because humans have low immunity to fight this virus. COVID-19 is a lung infection resulting from the SARS-CoV-2 virus (World Health Organization, n.d.). The virus's origin is still unidentified, but it is believed to have originated in Wuhan, China in December 2019. Since then, the disease has spread rapidly worldwide, becoming a global pandemic. The COVID-19 pandemic spread globally, devastatingly affecting health, socioeconomics, and the environment. Its impact resulted in economic and social changes and trends. The political role is given special consideration in this situation. COVID-19 pressures political structures worldwide, whether dictatorial or democratic, to provide adequate emergency epidemic control. Global leaders have taken critical actions to control the virus's spread. They have increased investment in healthcare infrastructure and resources and implemented social distancing measures, such as lockdowns and restrictions on public gatherings, to slow the spread of the virus. In addition, there has been a tendency to use technology for everyday tasks. For example, many companies use online shopping and delivery services, and schools adopt remote work and education activities to reduce the risk of exposure to the virus. The government has given guidelines for these solutions.

The pandemic has also affected political parties, with many elections and political events postponed or cancelled due to public health concerns. During this epidemic, the policy decisions taken by political leaders should be sufficient to solve the crisis and strengthen the confidence of the community and followers. Both positive and negative impacts of the COVID-19 epidemic have been observed within politics. Hence, it is essential to analyze people's ideas about the changes that have taken place in politics since the pandemic situation.

In the global community, opinion analysis and research on social media have become increasingly important topics among researchers because of the vast amount of information representing social opinion that can be easily accessed and specialized. Moreover, performing comparative analyzes on data from various countries may aid in making relevant global comparisons (Howoldt et al., 2023). Social media analysis is essential for politicians because it provides beneficial insights into public opinion and sentiment. Understanding public opinion, identifying key issues that are important to their constituents, targeting specific groups of voters based on their interests and behaviors, and measuring the effectiveness of their campaigns to help politicians respond quickly to crises and manage their reputations are some reasons that social media analysis is essential for politicians. To achieve this, Twitter has launched a newly developed track called the "Academic Research Product Track" (Parack, 2021). Twitter is a public platform that is visible to everyone. Therefore, it is a widely-used microblogging platform and a quick and straightforward platform to send short messages and share tweets, GIFs, article links, videos, and other media, and users can successfully spread political opinions and general information among their networks.

Accordingly, the main problem of this research is identified as "How can the topics that were discussed in relevance to political changes during the COVID-19 pandemic be found?" Based on this, the main objective of this study is to extract the topics discussed as most relevant to political changes during the COVID-19 pandemic using a topic modeling approach. This analysis helps to gain a deeper understanding of uncovering hidden themes of political changes. Fig. 1 shows the graphical representation of the mapping of research questions and objectives in this study.

We used topic modeling as a methodology to identify patterns of political behavior related to the identified research questions. Topic modeling is one of the natural language processing (NLP) techniques used to extract data and identify topics or themes in a collection of text documents (Manthiramoorthi, n.d.). It involves using algorithms to automatically identify patterns and clusters of words that frequently occur together in the text. The main objective of this research is to extract topics relevant to the political changes during the COVID-19 pandemic. This gives a better understanding of how political parties have adapted and responded to the pandemic in terms of their policies and actions. The results help to compare and contrast the political changes and responses to the pandemic across different regions and countries. The relevant parties can evaluate the effectiveness of different political strategies and responses to the pandemic based on public reactions and feedback on social media. They can apply those changes in future pandemic situations and make better political plans for their future campaigns in the country.

The study's remaining sections are structured as follows: The second section describes the literature review.

 https://doi.org/10.1633/JISTaP.2025.13.1.2

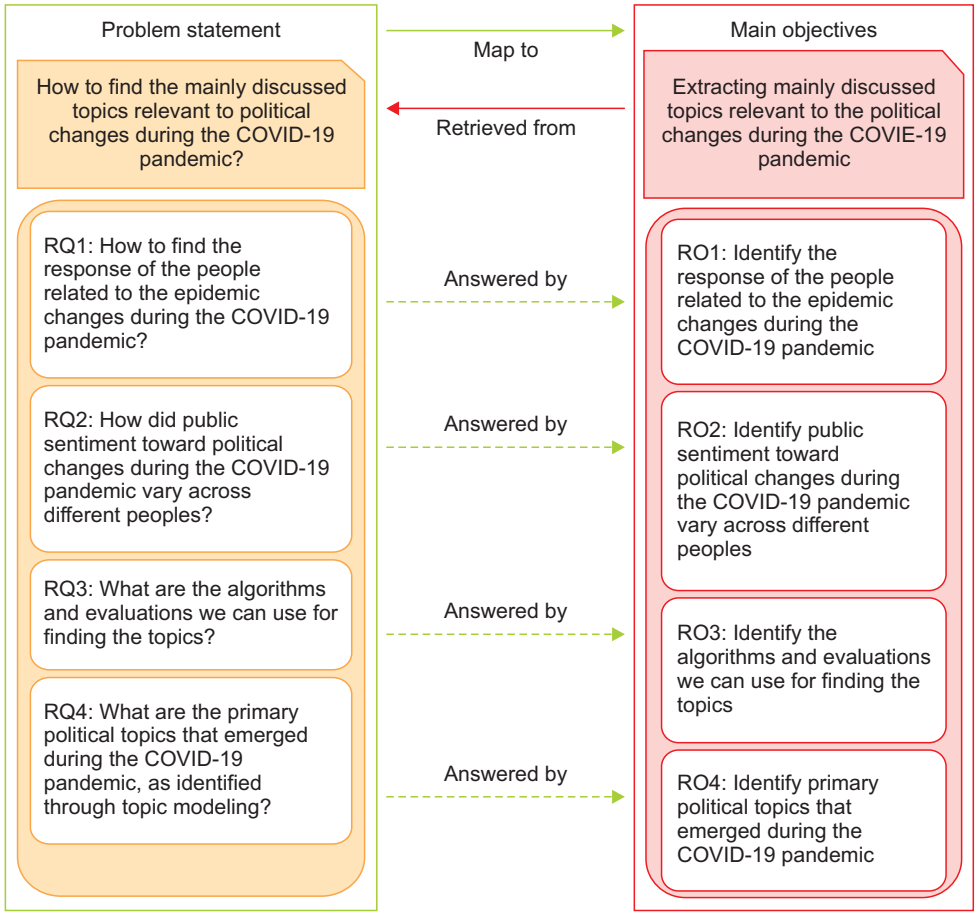


Fig. 1. Mapping of the research question and research objectives. COVID-19, coronavirus disease; RQ, research questions; RO, research objectives.

The third section explains the methodology. The fourth section explains the results and discussion. Finally, section five describes the conclusion of our research.

2. LITERATURE REVIEW

2.1. Social Media and Politics

Various research papers observe that media and political information or news influence public opinion widely. Because of this, social media has become an increasingly important topic among researchers. Social media analysis is essential for politicians because it provides valuable insights into public opinion tracking and measuring. Candidates create systems for monitoring and evaluating online activity, such as communication on Facebook, Instagram, and Twitter, and compare it with their conventional ones. Understanding public opinion, identifying key issues that are important to their constituents, targeting specific groups of voters based on their interests and behaviors, and measuring the effectiveness of their campaigns to help politicians respond quickly to crises and manage their reputations are some reasons social media analysis is essential for politicians (Mameli et al., 2022).

Meduru et al. (2017) realised that there is a connection gap between citizens and their government. As a result of this communication gap, policy decisions implemented directly or indirectly by the government may fail. Further research suggests using the Vader Lexicon to analyze opinions using Twitter feeds for political analysis. The use of Twitter to frame the COVID-19 virus was investigated using four methods by leaders of the "Five Eyes" intelligence-sharing group, which includes the United States, the United Kingdom (UK), Australia, New Zealand, and Canada. It has been suggested that implementing a new algorithm utilizing the benefits of several algorithms can increase the accuracy of the algorithms so that they can be effectively used in forecasting.

A sentiment-based classifier using machine learning and a "Fuzzy Set and Rough Set approach" for predicting the results of the 2019 presidential election of Sri Lanka, using tweets from September 1 to November 10, 2019 (two months before the election) for both "Gotabaya Ra-

japaksa" and "Sajith Premadasa," was also developed by Perera and Karunanayaka (2022). This study reveals flaws and limits in sentiment-based classifiers regarding political tweets associated with this 2019 presidential election case study. It has been suggested that implementing a new algorithm utilizing the benefits of several algorithms can increase their accuracy, so that they can be effectively used in forecasting. Howard and Parks (2012) suggest that communication researchers are uniquely positioned to analyze the relationship between social media and political transformation in terms of both causes and effects. In Bermingham and Smeaton (2011), when considering the Irish General Election, the case study to investigate the potential of modelling political sentiment through social media data mining, discovered that social analytics using both volume-based measures and sentiment analysis is predictive. Several observations are related to monitoring public sentiment during an election campaign, such as examining a variety of sample sizes, periods, and methods of qualitatively exploring the underlying factors. This could involves investigating how politicians' mistakes on social media platforms affected people's opinions of Twitter communication and their responses to other politicians' tweets (Lee et al., 2020). These results highlighted that political gaffes made by politicians on social media may, ironically, improve the success of their efforts by creating a sense of genuine dialogue.

In a study of presidential elections in Egypt, users from different cultures and backgrounds on social networks such as Facebook and Twitter focused on a large number of textual comments that reflected their views in various situations (Elghazaly et al., 2016). The study used support vector machine (SVM) and Nave Bayesian to classify the Arabic text. The comparison showed that the Nave Bayesian technique has the best accuracy and lowest error rate. The study by Bose et al. (2019) further demonstrate how data twitter can be used and how the national research council Emotion Lexicon can be used to analyze politics. According to the 14th Gujarat Legislative Assembly elections of 2017 in India, a dataset of Twitter messages related to the election was used to predict the chances of party victory. Then, the sentiment was analyzed as positive, negative, and neutral, and the results were summarized.

2.2. COVID-19 Pandemic and Politics

The COVID-19 pandemic has had a significant impact on politics around the world. In some countries, the pandemic has led to political instability and unrest, mainly because governments have been slow to respond to the crisis, and the lack of management and information received by the health sector has led to social unrest. Summarizing these facts, politics has been tested through several variable aspects during the epidemic period. Palmer (2021) investigates the political attitudes and behavior of residents of states in the United States. The data set consists of demographic factors on education, race, age, and social trust, with tested policy specifications, social trust, and relationship with policy outcomes. Statistical diagnostic methods were used to examine outliers' distribution and the possible influence and impact to uncover essential failings within all models. The analyzes' assumptions have been meticulously interpreted using data.

Rieger and Wang (2022) examined data from 57 countries between March 20 and April 22, 2020 on how various factors such as education, actual stringency level, media freedom, and death rate affect judgments of government policy responsiveness and how people perceive government responses. Survey evidence shows that conspiracy theorists tend to take government counterattacks to be harsh. Purushothaman and Moolakkattu (2021) look at previous studies, press reports, and press releases by government agencies to gather the necessary data from India. The research questions investigated in their article are government response to the pandemic, containment measures, government consolidation of power and international status-seeking, and foreign policy related to the impact of the pandemic and vaccination diplomacy. Montiel et al. (2021) have mapped the political rhetoric of national leaders and characterized global changes during the COVID-19 pandemic. For this, the main rhetorical stories in the publicly available speeches of 26 countries were used as a sample. Lipscy (2020) proposes a broad investigation centered on causes, reactions, and transformations. The COVID-19 pandemic hence highlights the need for more scholarly attention to the political crisis. International relations have focused on military crises, neglecting non-militarized crises such as pandemics. To understand crises, researchers must examine the threat, uncertainty, and time pressure that impact political and economic outcomes. Another paper discusses predicting responses using political-related blogs as primary data. Yano et al., (2009) examined 40 blog sites from November 2007 to October 2008, and the latent Dirichlet allocation (LDA) topic model was used to generate the topics. This provides a qualitative discussion model on a novel comment prediction task, which uses models to predict which blog users will comment on a given post.

2.3. Methodological Review

Different methods are utilized for evaluating social media data and political facts. To evaluate public opinion on the central UK government during the first phase of the COVID-19 epidemic, structural topic modeling was used to extract topics from 4,000 free-text survey responses collected between October 14 and November 26, 2020 (Wright et al., 2022). Overall, they classified answers as positive or negative using sentiment analysis. Specifically, Bogdanowicz and Guan (2022) identified the 12 most popular topics in our dataset from April 3 to April 13, 2020, such as politics, health care, community, and economy, and discussed their growth and changes over time. While macro-level growth has been experienced over time, it is clear that micro-level changes in topic composition have also been exhibited. The research by Rathnayaka and Rupasingha (2023) outlined the sentiment analysis on Twitter data, focusing on the COVID-19 pandemic and political changes' impacts on people's lives. Their work employed both machine learning and deep learning techniques, specifically SVM, artificial neural network, and long short-term memory for the categorization of tweets into positive, negative, and neutral sentiments.

Guo et al. (2016) examine two big data text analysis methods: lexicon-based analysis, the most popular automated analysis approach in social science research, and unsupervised topic modeling (LDA analysis). Analyzing 77 million tweets about the 2012 United States presidential election, the study provides a starting point for scholars to evaluate the effectiveness and validity of journalism and various computer-based methods. Notably, the study by Guo et al. (2016) revealed that the two methodologies significantly produced different results. The LDA-based analysis, in particular, was able to read more tweets and reveal deeper subtleties of the conversation. The LDA-based analysis was also determined to be more valid than the dictionary-based technique based on human coder ratings. The LDA-based analysis is cost-effective since it requires little human labor. This demonstrates that, in general, LDA-based analysis outperforms other computer-based approaches.

Olbrich and Yamshchikov (2018) have chosen probabilistic LDA topic modeling to develop political space, which is widely used in NLP, with a particular focus on developing unsupervised text analysis methods. As LDA can generate meaningful political spaces, new dimensions of political discourse can be identified. Because of the qualitative nature of working with textual data, it is evident that the evaluation of the LDA model's findings is both quantitative and qualitative. This paper introduces probabilistic latent semantic analysis (PLSA) and LDA as Twitter user opinion extraction methods. It is unrelated to the COVID-19 pandemic and politics, but provides a comprehensive explanation regarding topic formatting (Gandhi, 2021). It utilizes Dirichlet priors for the document-topic and word-topic distributions, allowing for better generalization.

Probabilistic topic models have recently received much attention because of their performance in various text-mining tasks such as retrieval, summarization, categorization, and clustering (Lu et al., 2011). In recent years, probabilistic topic modeling has been frequently applied to research topic modeling to focus on finding hidden topics in large documents.

The comparison summary with previous studies is discussed in Table 1 (Bermingham & Smeaton, 2011; Bose et al., 2019; Elghazaly et al., 2016; Lee et al., 2020; Montiel et al., 2021; Palmer, 2021; Purushothaman & Moolakkattu, 2021; Rathnayaka & Rupasingha, 2023; Rieger & Wang, 2022; Wright et al., 2022; Yano et al., 2009).

2.4. Research Gap

Considering the COVID-19 pandemic, we determine the most frequently discussed political themes on Twitter. The critical point is that we could not find any studies focused on the COVID-19 pandemic and political changes that applied a topic modeling approach to extract the relevant topics. The proposed approach is mainly targeted to fill this research gap. Some other relevant research is limited to one country and uses relatively small data sets. When considering the source of data acquisition, some studies collected data from weblogs, Google forms, and polls conducted in different countries. According to the literature, using public opinion platforms such as social media to get accurate ideas is better. It provides opportunities for learning about and comprehending people, their potential, hobbies, and viewpoints.

Due to the ease of obtaining and accessing data, attention has been drawn to Twitter. Hence, the study's main focus was on the analysis of tweets using standard topic modeling methods. It is capable of correctly studying changes in politics during the COVID-19 epidemic. The literature-identified LDA and PLSA topic models were more efficient and accurate for the proposed approach. As a result, the entire globe has become the subject of research, obtaining data far more extensively than with prior data collection methods. LDA and PLSA algorithms are projected to discover these study fields.

Table 1. Summary of comparison with previous studies

Reference	Data	Methodology	Objective	Limitation/difference
Bermingham & Smeaton (2011)	Twitter data	Sentiment analysis through SVM algorithm	Irish general election as a case study for investigating the potential to model political sentiment through Twitter data	Targeted only the Irish general election and limited to one nation; only conducted sentiment analysis
Lee et al. (2020)	Total of 450 adults via email through an online survey in South Korea	Statistical analysis	How politicians' Twitter blunders might affect individuals' beliefs about the nature of Twitter communication, and subsequently, their reactions to other politicians' tweets	Only conducted a statistical analysis; limited to one country
Rathnayaka & Rupasingha (2023)	Twitter data	SVM, ANN, and LSTM algorithms	Sentiment analysis on twitter data relevant to politics during COVID-19	Only conducted sentiment analysis
Elghazaly et al. (2016)	Twitter data targeting presidential elections in Egypt	SVM and Naïve Bayesian algorithm	Political sentiment analysis using Twitter data	Limited to one election-related data and country; only conducted sentiment analysis
Bose et al. (2019)	Twitter messages related to recent 14th Gujarat Legislative Assembly Election	NRC Emotion Lexicon, ParallelDots AI APIs	Predicting the chances of winning party by utilizing public opinion	A few numbers of data used (1,000); only conducted an emotion analysis; limited to one country
Palmer (2021)	Through survey	Statistical models (regression analysis)	Measuring the political attitudes and behavior in a pandemic: factors affecting compliance with COVID-19 policies	Only conducted a statistical analysis; a few number of data used
Rieger & Wang (2022)	Online survey	Statistical analysis (structural equation model)	Study people's perceptions of government reactions during the pandemic	Only conducted a statistical analysis
Purushothaman & Moolakkattu (2021)	Press reports and press releases by government agencies	A descriptive and analytical approach	Analysis of the politics of the COVID-19 pandemic in India	Only conducted a statistical analysis; limited to one country
Montiel et al. (2021)	Publicly available speeches	Principal component analysis, LDA algorithm	Maps political rhetoric by national leaders during the COVID-19 pandemic	Research depends on the public speeches of the politicians
Wright et al. (2022)	Survey response data	LDA algorithm	Analyzing public opinion about the UK government during COVID-19	Limited to one country
Yano et al. (2009)	Online political blog post	LDA algorithm	Predicting response to political blog posts with topic models	A few numbers of data; not relevant to the pandemic situation

SVM, support vector machine; ANN, artificial neural network; LSTM, long short-term memory; COVID-19, coronavirus disease; NRC, national research council; AI, artificial intelligence; API, application programming interface; LDA, latent Dirichlet allocation; UK, United Kingdom.

3. METHODOLOGY

The primary objective of this research is to compare topic modeling strategies with their outcomes. We examine the research strategy, data collecting and pre-processing, feature extraction, and implementation of the data analysis method as part of the research methodology by topic modeling approach. Finally, topics are analyzed using two topic modeling approaches, LDA and PLSA. Fig. 2 explains the architecture of the proposed approach.

 https://doi.org/10.1633/JISTaP.2025.13.1.2

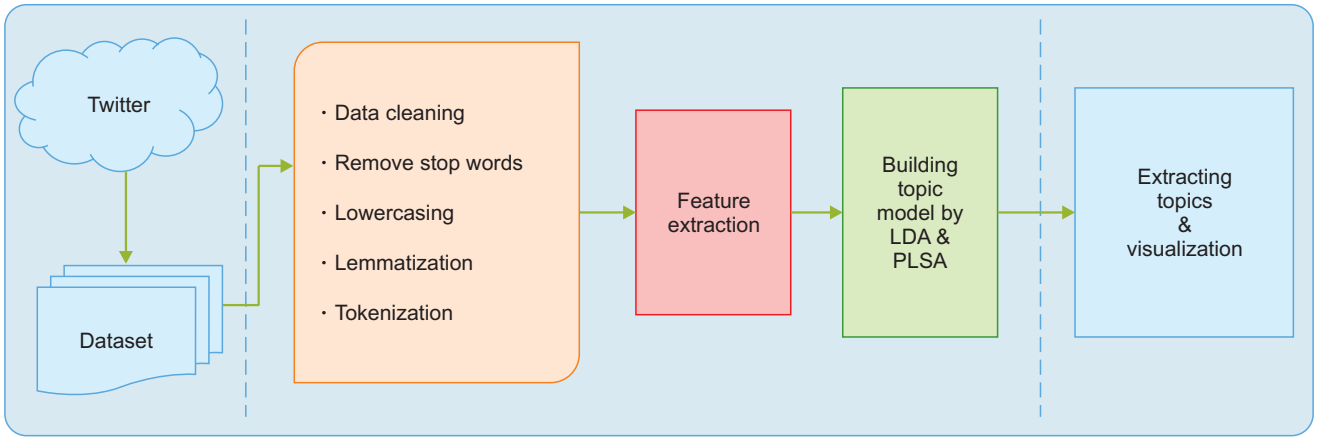


Fig. 2. Architecture of the proposed approach. LDA, latent Dirichlet allocation; PLSA, probabilistic latent semantic analysis.

	tweet
2	Donâ€™t befriend people based on political opinions because the moment that opinion changes, the friendship ceases to exist.
3	I already voted, but everyone else must have the opportunity to vote from home! #CenterForVoterInformation many thanks for your efforts. We need political changes so that people donâ€™
4	@itzDanzi There are going to be so many social, political, economic, scientific & environmental changes that will stem from all of this. The extent, impact & sustainability of them
5	@Faazzla @GNedland97 @peterdaou I agree, unfortunately that's the reality of our political situation.Big far reaching legislative changes are driven by the presidential cycle, and by new lea
6	@WTingzz @BeeBabs @msangelinejolie Either way problem will never go away until black people as a collective realise that twitter supremacy doesnâ€™t translate into real life. We need
7	COVID-19 Case Fatality Rate â€"May Be Considerably Less Than 1%â€™ â€“ Dr. Anthony Fauci [beware, Fauci changes with changing political winds] https://t.co/IhlEMXtM1c
8	@PersonRespector @Patrick_Wyman This. Iâ€™m very interested in the lasting disruptions (political, social, demographic) etc, where the pandemic is truly the catalyst for these changes.
9	@AndreaJ19954134 @liliebayer No, formally there have been no changes in the number of MPâ€™s. The Swedish political parties have unanimously agreed to limit the number of MPâ€™s t
10	@LifeSciVC My guess is that social/political changes after the pandemic will have way more impact than technology changes. Like Medicare for all might actually be possible and save many
11	@chrisdeerin Political environment as important as the science. Changes in the temperature of that environment precede policy alterations. Fascinating how the Scandinavian countries diffe
12	@LiberlHnyBadger @cenkuygur @AOC @politico There is NEVER a time to play nice w/a privately-run corrupt elitist political career club that makes & changes its rules at whim - threat
13	@realDonaldTrump I like things on both parties and dislike other things. We never get anything accomplished bc as soon as power changes at Pres/House/Senate etc, what was applied gets
14	@KCRG The political criticism here is insufferable. I've personally spoken to PPE purchasing managers at hospitals late last week. They've noted their distribution / supply changes on daily b
15	@mamarocks54 Its so odd. I have picked a political but when they are wrong or I didn't agree I certainly would say so. Why are his fans so odd the way they back everything he says, Even wh
16	@Dannythefink If just 1:50 citizens changes their behaviour because they read and took on the stay at home message, then, over 1 million people will have changed. I note a lot of the compl
17	The Report included individual success stories that demonstrate that we do possess the knowledge and tools to improve this situation â€" provided that we have the political will to enact sy
18	@WhiteHouse @realDonaldTrump This is what a responsible government official says:Gov. Ralph Northam of Virginia on his decision to issue a stay-at-home order for the state. Heâ€™s ma
19	CALL: Australiaâ€™s leading scientists call for data underpinning COVID-19 decisions to be made public https://t.co/m9mLvkqLK7 "open publication of data & evidence supporting gove
20	#Kantar Canada has released results from our G7 study of public reactions to COVID-19. Although most Canadians are concerned about personal impact, trust in Government for information

Fig. 3. A portion of the data set.

3.1. Data Collection

Twitter social media was used to collect data relevant to the world's political changes during the COVID-19 pandemic. Tweets were collected during March 2020 and December 2021, the peak period of the COVID-19 pandemic, and 10,658 tweets were collected through the Twitter application programming interface. First, a Twitter application development account was registered to scratch tweets from Twitter. In the Postman workspace, we made a fork through workspace and sent a fetch request to search the entire archive by providing values for several variables (query=COVID-19, political changes, date=20200301000, toDate=20200630000, and maxResults=100). Comma separated values formatting was used in the data set. Fig. 3 shows a portion of the data set. The keywords used are "COVID-19," "Political changes," and "Government decisions." Data were downloaded in JavaScript Object Notation format by sending a fetch request. Our keywords were chosen to reflect a global focus rather than being specific to any particular country or region. However, the collection process was primarily focused on English-language tweets.

3.2. Data Pre-Process

Data pre-processing is modifying or removing data before use to ensure or improve performance. This phase of the data mining process is essential. The data was pre-processed using Python libraries with NLP techniques to prepare the data suitable for the analysis. First, we import some required libraries such as "re, pandas, numpy, gensim, spaCy, pyLDAvis." These modules are available for performing text preprocessing, topic modeling, visu-

alisation, and plotting tasks in the subsequent code. For example, spaCy and Natural Language Toolkit (NLTK) are NLP libraries that provide various functionalities, such as tokenization, part-of-speech (POS) tagging, and dependency parsing. Gensim is a popular Python library for topic modeling analysis. The NumPy library provides support for mathematical operations on arrays and matrices. pyLDAvis is used for visualization. After this, the following mandatory steps are proceeding.

3.2.1. Data Cleaning

This step aims to clean and normalize the text data in the "tweet" column by performing several pre-processing steps. Also, as this study considers tweets in the English language only, therefore we installed the "spaCy" NLP library that loaded a pre-trained English language model called "en_core_web_sm" and removed other language tweets from the data set.

3.2.2. Prepare the Stop Words

The NLTK library can retrieve a list of English stop words using the "stopwords.words('english')" function call. It provides a collection of typical English terms. Stop words consist of pronouns and articles, including "the," "is," and "and." The "stop_words.extend([...])" statement is used to add additional stop words to the existing list.

3.2.3. Lowercasing

This step is generally taken to strengthen the stability of text analysis. From Python, "text.lower()" converts all the uppercase letters in the text to lowercase.

3.2.4. Lemmatization

Lemmatization is the process of reducing words to their base or root forms (lemmas) to normalize the text. It involves removing the inflectional endings from the words and returning the base form of the word. It groups different forms of words as one single form of the word. By default, it keeps only nouns, adjectives, verbs, and adverbs. It uses the spaCy library to tokenize each sentence in the texts and then lemmatizes each token based on its POS tag ("allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV']").

3.2.5. Tokenization

Tokenization is the process of obtaining a token from tweets. At this point, a longer paragraph is converted to tokens. The example shows individual words separated by spaces in a text document, which are referred to as tokens. The NLTK library has a Tweet Tokenize module that can be used to extract tokens from tweets where all words are punctuated. Sentences are pre-processed using the "gensim.utils.simple_preprocess" function before being entered into the "sent_to_words" function. The "deacc=True" argument is given to the simple pre-process method to ensure that punctuation marks are removed from the tokens. The following Table 2 explains the example tweet for the above pre-processing steps.

3.3. Feature Extraction

Feature extraction in the LDA topic model is the process of identifying the most important words or phrases for a better understanding of the context (Blei et al., 2003). After cleaning and normalizing the data, we need to con-

Table 2. Example for pre-processing steps

	Example
Tweet before the preprocess	@WhiteHouse @realDonaldTrump This is what a responsible government official says: Gov. Ralph Northam of Virginia on his decision to issue a stay-at-home order for the state. Heâ€™s making his decisions on the response to Covid-19 based on science and data
Data cleaning	WhiteHouse realDonaldTrump This is what a responsible government official says Gov Ralph Northam of Virginia on his decision to issue a stay at home order for the state making his decisions on the response to COVID 19 based on science and data
Prepare the stop words	WhiteHouse realDonaldTrump This what responsible government official says Gov Ralph Northam Virginia his decision issue stay home order state making his decisions response COVID 19 based science data
Lowercase	whitehouse realdonaldtrump this what responsible government official says gov ralph northam virginia his decision issue stay home order state making his decisions response covid 19 based science data
Lemmatization	whitehouse realdonaldtrump responsible government official say decision issue stay home order state make decision response covid base science data
Tokenization	["whitehouse" "realdonaldtrump" "responsible" "government" "official" 'say" "decision" "issue" "stay" "home" "order" "state" "make" "decision" "response" "covid" "base" "science" "data"]

 https://doi.org/10.1633/JISTaP.2025.13.1.2

vert it into numerical representations for modeling: that is, converting unstructured data into a structured numerical format. This study used Bow to convert into a numerical representation. The bag-of-words model is a way of representing modeling text with algorithms. The code "[id2word.doc2bow (text) for text in texts]" represents this process. The "doc2bow" counts the occurrence frequencies of each distinct word, and "id2word" maps the word_id to a unique word in the corpus using the Gensim library.

3.4. Applying Topic Models

To identify the key themes and topics about political behavior that occurred during the COVID-19 pandemic, this study used LDA and PLSA topic modeling analysis. PLSA and LDA are both topic modeling algorithms used to analyze large collections of documents. The analysis will highlight the most significant changes in political speech as well as people's opinions on these changes. Here, we examine which method of analysis provided the most accurate information.

The existing techniques, LDA and PLSA, were used to obtain more accurate results effectively and efficiently. The literature-identified LDA and PLSA topic models were more efficient and accurate than other methods for the proposed approach. They are successful methods, and each document is viewed as a mix of multiple distinct topics. The advantage of the LDA technique is that one does not have to know in advance what the topics will look like. LDA automatically provides topics based on patterns in the data. PLSA provides a clear probabilistic interpretation of the relationship between documents and topics. The proposed approach primarily relies on such established techniques such as LDA and PLSA when considering these advantages.

3.4.1. LDA Topic Modeling

LDA is a topic modeling technique that represents the relationship between multiple documents in a corpus. This technique allows a set of observations to be explained by unobserved groups that define why some parts of the data are similar (Alamoodi et al., 2022). It uses a Bayesian approach and assumes a Dirichlet prior over the topic distributions. Fig. 4 shows the algorithm.

The variable names are defined as,

M is the number of documents in the corpus.

N is the number of words in the document.

K is the number of topics.

α is Dirichlet parameter (controls per-document topic distribution).

θ_m is document topic distribution.

$Z_{m,n}$ is word topic assignment.

$W_{m,n}$ is observed word.

φ_k is topic-word distribution.

β is Dirichlet parameter (controls per-topic word distribution).

LDA estimates the proportion of words in a document assigned to a certain topic, and then finds the fraction of words assigned to a topic across all texts. These themes are then qualitatively examined to validate their content relevance.

3.4.2. PLSA Topic Modeling

PLSA is a generative model that assumes each document is formed from a mixture of latent themes. Each word in a document is generated with a specific probability by one of these topics. PLSA aims to estimate the probability distributions for each document's subjects and terms. It is a topic modeling technique for data analysis

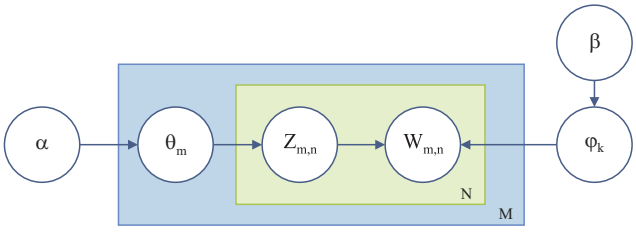


Fig. 4. Latent Dirichlet allocation algorithm. M, number of documents in the corpus; N, number of words in the document; α, Dirichlet parameter (controls per-document topic distribution); θ_m, document topic distribution; $Z_{m,n}$, word topic assignment; $W_{m,n}$, observed word; φ_k, topic-word distribution; β, Dirichlet parameter (controls per-topic word distribution).

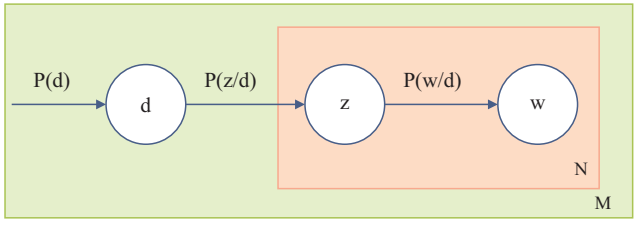


Fig. 5. Probabilistic latent semantic analysis algorithm. d, number of documents in the corpus; w, number of words in the document; z, number of topics; P(d), probability of selecting a specific document (d) from the entire document collection; P(z/d), topic (z) is present in the selected document (d) with probability; P(w/d) is word (w) is drawn from the topic (z) with probability in a given topic.

with applications in information retrieval and filtering, NLP, machine learning from text, and other fields. The PLSA algorithm generated hidden topics based on the importance of words and the similarity of topics in a document (Khotimah & Sarno, 2019). Fig. 5 shows the algorithm.

The variable names are defined as,
d is number of documents in the corpus.
w is number of words in the document.
z is number of topics.
P(z/d) is topic (*z*) is present in the selected document (*d*) with probability.
P(w/d) is word (*w*) is drawn from the topic (*z*) with probability in a given topic.

4. RESULTS AND DISCUSSION

Here, we discuss the extracted topics relevant to the political changes in the COVID-19 pandemic using topic modeling techniques. This evaluation used Microsoft Windows 10 Pro (Microsoft Corp. Redmond, WA, USA) on a personal computer with an Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz, 4 Core(s) Processor, and 8GB RAM. Python was used as a programming language to implement topic modeling algorithms.

4.1. LDA Visualization Analysis

The LDA visualization analysis used the "LDAvis" R package (Sievert & Shirley, 2014) for the visualization, which is intended to explore LDA topic models by varying the weights of top-ranked topic words. Inter-topic distance map is a web-based interactive visual representation of topics computed using LDA. This is the most often used visualizing library.

The topics are plotted on an inter-topic distance map by default; the x and y axes here are labeled "PC1" and "PC2" (principal components 1 and 2). Fig. 6 shows that 15 topics were chosen, and bubbles represented the topics with their frequencies. The size of each bubble indicates the frequency of that topic, and the distance between the bubbles reflects the semantic similarity between topics. The topic format will be successful if it uses large, evenly-spaced bubbles that do not overlap. Fig. 6 shows that topics 2, 7, 9, and 15 are clustered independently, and others overlap event topics. In considering overlapping topics, some keywords are commonly used for each overlapping topics. For example, "covid" and "government" keywords are common to topics no. 10, 11, and 14. However, those clusters cover specific topics, and they are clusters of mutually distinct topics.

Fig. 7 shows the 30 most relevant terms and ranks them in order of relevance from top to bottom and fills the bar chart. Considering the keywords shown in LDAvis, the words with the highest frequency are "good," "decisions," "make," "look," "covid," "trackle," "people," "inquiry,"

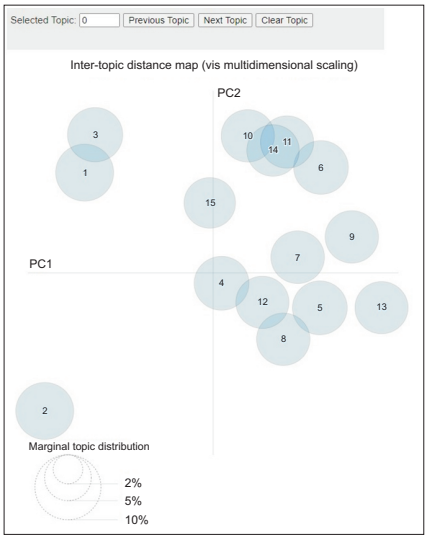


Fig. 6. Inter-topic distance map (topic bubbles).

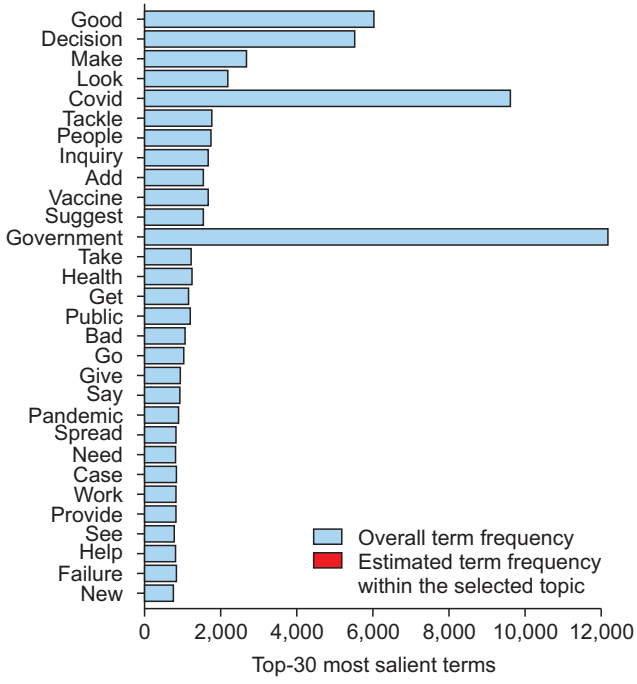


Fig. 7. Most salient terms.

 https://doi.org/10.1633/JISTaP.2025.13.1.2

"add," and "vaccine."

Fig. 8 (Sievert & Shirley, 2014) shows the interaction of left and right sides in the inter-topic distance map. For example, when topic number 10 is selected in the right panel, the left panel terms with their frequencies are represented by red bars. Primarily, LDAvis enables one to choose a topic and display the most relevant keywords for that topic. Relevance is defined by λ, which is the weight assigned to the probability of the term in a topic relative to its lift. The value is intended between 0 and 1. When λ=1, terms are ranked according to their probabilities inside the topic, whereas when λ=0, terms are re-ranked solely according to their lift; Sievert and Shirley (2014) realized that λ=0.6 was the optimum interpretable result.

4.2. LDA Topic Analysis

LDA assumes that each document is composed of a distribution of topics, and each topic is composed of a distribution of words. The task of this LDA algorithm is to build the hidden topic ("latent"). Table 3 presents the topics and keywords of tweets in each topic produced from 10,792 tweets. When considering Table 3, the study identified the five key themes and 15 topics about political behavior that occurred during the COVID-19 pandemic. "No" in Table 3 represents the relevant cluster numbers in the Inter-topic distance map in Fig. 6.

As shown in the following Table 3, we defined and interpreted the five theme findings. These five themes are defined based on the results of the inter-topic distance map in Fig. 8 (Sievert & Shirley, 2014).

In Fig. 8 (Sievert & Shirley, 2014),
- 15 circles on the left side - 15 extracted topics from the LDA model.
- 30 terms on the right side - Set 30 keywords for the selected topic. Here we selected Topic 10 (highlighted in red color).

Fig. 8 (Sievert & Shirley, 2014) visualizes the relevant set of 30 keywords for each topic. Among the 30 keywords, the lengthy red bars at the top show the top keywords for the selected topic [in Fig. 8 (Sievert & Shirley, 2014), selected topic 10]. Based on these top keywords, we manually decided on the name for that topic based on

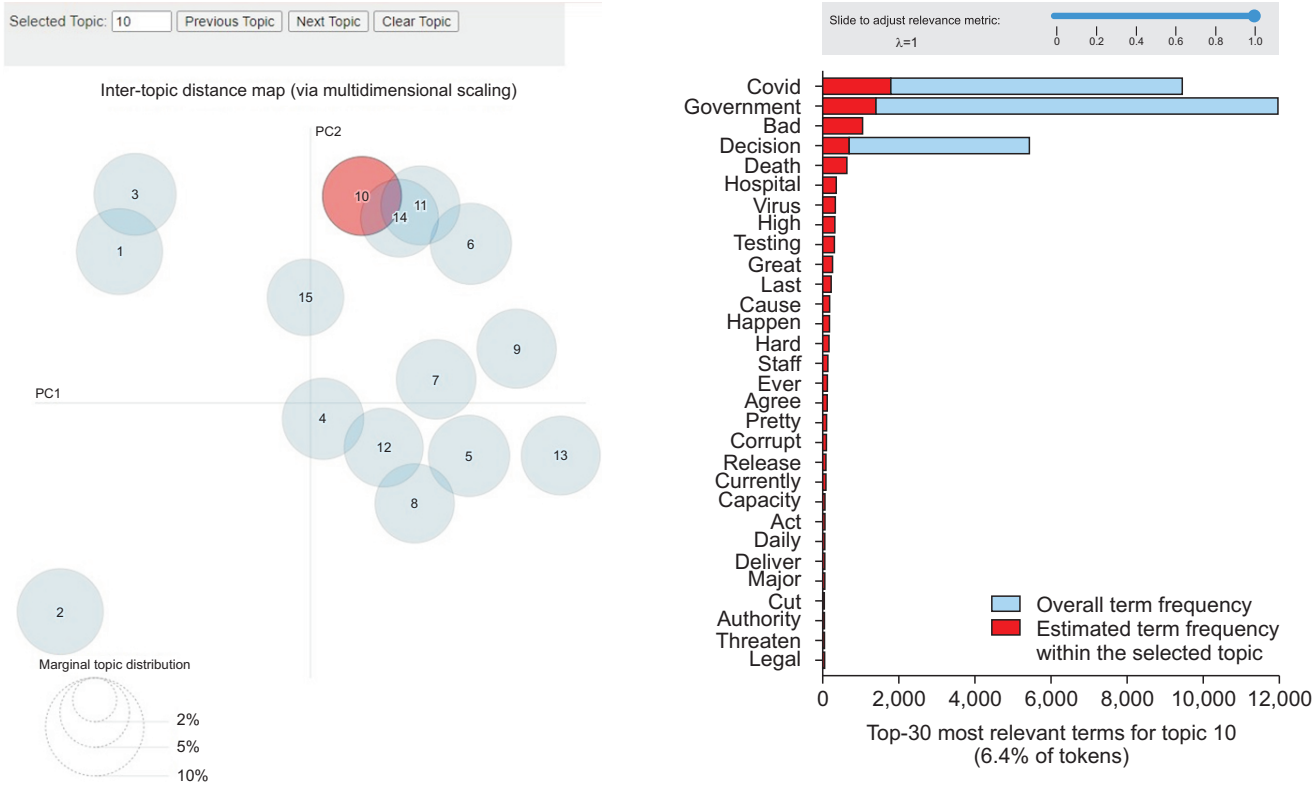


Fig. 8. How the inter-topic distance map is displayed when the topic is selected. 1. Saliency(term w)=frequency(w)×[sum_t p(t|w)×log(p(t|w)/p(t))] for topics t. 2. Relevance(term w|topic t)=λ×p(w|t)+(1-λ)×p(w|t)/p(w); data from the article of Sievert & Shirley (2014).

Table 3. Topics and themes generated through LDA

No.	Topic	Keywords	Example Tweet
Theme 1: Government Responses for Managing the COVID-19 Pandemic			
1	The travel mandates	"decision," "government," "follow," "information," "let," "mandate," "travel," "allow," "guideline"	"Updated information on Covid-19 vaccine mandates"
4	Ensuring public safety and the federal government	"good," "government," "state," "federal," "safe," "require, "guideline', "meet," "covid"	"With the federal government handing off decisions, states have been the primary drivers behind moves with the most impact on the pandemic's spread"
11	Enforcing regulations to manage the spreading of the virus	"government," "covid," "spread," "stop," "protect," "rule," "prevent," "community," "responsible," "disease," "affect"	"A citizen while we appreciate the government for trying very hard to stop the spread of Covid-19"
14	Efforts to control the spreading	"government," "get," "covid," "measure," "come," "number," "big," "result," "suffer," "handle"	"Statement on the provincial government's announcement today regarding additional Covid-19 public health measures"
Theme 2: Government decisions for COVID-19			
3	The decision to take control spreading of the virus	"decision," "government," "take," "level," "control," "never," "close," "cover," "decide"	"Governments on all levels are either unable to make the necessary decisions needed to control Covid-19"
6	Real-time monitoring of the situation	"covid," "suggest," "government," "pandemic," "case," "watch," "coronavirus," "important," "action," "priority"	"The well worth watching particularly if you are questioning Government decisions and Covid-19 cases"
10	Testing and hospital capacity	"covid," "government," "bad,", "death," "hospital," "virus," "high," "testing," "staff," "capacity," "daily," "threaten"	"As if the first two waves of Covid-19 hospitalizations in the USA weren't enough to inspire serious political changes to stop the coronavirus"
Theme 3: Public Response to Government Measures for COVID-19			
5	COVID-19 testing and public response	"provide," "help," "test," "report," "response," "covid," "policy," "datum," "low"	"All our winter events will continue as planned, Covid-19 testing government guideline s and the Good To Go Covid-19 industry standard"
7	Issues, changes, and guidelines on public health	"change," "regard," "covid," "issue," "guidance," "government," "update," "stay," "social," "post"	"Warns some against taking the vaccine. The key issue now is our regime restrictions tight enough to reduce the spread of infection. Some very tough decisions needed"
Theme 4: Social Influence			
8	A new lifestyle with pandemic	"go," "think," "time," "job," "thing," "life," "start," "really," "live," "try"	"It is absolutely true that Covid-19 has changed the lifestyle of many people all over the world. Along with human beings, animals suffer a lot due to this virus"
9	Economic changes	"covid," "country," "effect," "business," "support," "lockdown," "idea," "restriction," "economic," "strategy"	"As a result of the Covid-19 lockdowns, numerous social and economic changes were intended! Consider the upward transfer of wealth and market share to Big Business since the start of the Covid-19 pandemic"
12	Leadership, media influence on increasing public health	"people," "need," "home," "care," "leader," "news," "point," "question," "risk," "mask," "wear"	"CTV NEWS: Civil society leaders call for evidence-based decisions for public health"
15	Effect of the education section	"tackle," "government," "see," "new," "school," "student," "rate," "place," "learn," "reopen"	"Working fewer hours, concentrating on things we need more, E-learning, most will discover their potential and talents"

 https://doi.org/10.1633/JISTaP.2025.13.1.2

Table 3. Continued

No.	Topic	Keywords	Example Tweet
Theme 5: Vaccination			
2	Global vaccine demand, failures, and distribution inequity	"inquiry," "vaccine," "health," "public," "failure," "demand," "repercussion," "globally," "inequity," "service," "immediate," "safety"	"Join us in demanding that the Covid-19 inquiry looks at the government's failure to tackle vaccine inequity globally, and the repercussions of these decisions on public health! because very immediate need to respond to the pandemic"
13	Role of experts and politics vaccination	"covid," "give," "vaccination," "expert," "political," "share," "medical," "vaccinate," "dose," "critical," "mandatory," "relate"	"Government isn't going to listen to an expert's opinions about Covid-19 and vaccination!"

LDA, latent Dirichlet allocation; COVID-19, coronavirus disease.

expert knowledge and feedback.

As shown in Fig. 8 (Sievert & Shirley, 2014), we extracted 15 topics, and considering all the extracted 15 topics, we divided them into five themes manually based on expert knowledge and feedback. The decided themes are 'Government Responses for Managing the COVID-19 Pandemic,' 'Government Decisions for COVID-19,' 'Public Response to Government Measures for COVID-19,' 'Social Influence,' and 'Vaccination.' Considering the visualization of LDA and PLSA, as well as feedback from experts and knowledge from relevant authorities and government agencies, we decided that the themes are accurate and meaningful.

4.2.1. Theme 1: Government Responses for Managing the COVID-19 Pandemic

The first theme began with the government's response to managing the COVID-19 pandemic. It focuses on the various measures, strategies, and actions that governments around the world have taken to address the challenges posed by the COVID-19 pandemic. This theme focuses on the impact of the following four topics.

4.2.1.1. Travel Mandates.

During the pandemic, travel mandates were one of the successful methods used to minimize the spreading of the virus. Quarantine requirements, mandates to obtain vaccination doses, lockdowns and shutdowns of countries, and foreign travel restrictions are the methods governments use to manage the movement of people and reduce the spread of infectious diseases.

4.2.1.2. Ensuring Public Safety and the Federal Government.

Public safety and the federal government were significant in that time period. The federal government is a general government concept practised in some countries where different powers are divided at different levels. Accordingly, it refers to the government's process of establishing public health.

4.2.1.3. Enforcing Regulations to Manage the Spreading of the Virus.

The COVID-19 pandemic has significantly affected society worldwide, requiring governments to take proactive measures to minimize the spread of the virus.

4.2.1.4. Efforts to Control Spreading.

Government actions and measures are important in addressing the global crisis of the COVID-19 pandemic. The implementation of public health strategies, data tracking and reporting, and effective response to emergencies can be identified as part of governments' efforts.

4.2.2. Theme 2: Government Decisions for COVID-19

Governments are entrusted with making decisions that affect the welfare and stability of a nation. This theme delves into the specific policy decisions made by governments in response to the COVID-19 pandemic. As part of this, we discussed how governments determine their strategy for COVID-19 testing and hospital capacity. Governments must stay updated about the state of the pandemic and adapt their strategies accordingly. This theme focuses on the impact of the following three topics.

4.2.2.1. The Decision to Take Control of the Spreading of the Virus.

Governments on all levels must be able to make the necessary decisions needed to control COVID-19. Authorities need to be effective in making decisions; otherwise, it will cause public displeasure.

4.2.2.2. Real-Time Monitoring of the Situation.

The word "today" indicates the need for real-time monitoring and rapid response to the evolving situation. Governments must stay updated about the state of the pandemic and adapt their strategies accordingly.

4.2.2.3. Testing and Hospital Capacity.

This theme focuses on a government's decisions regarding COVID-19 testing and hospital capacity. Testing can be helpful to minimize the risk of spreading COVID-19 to others. There are two main types of viral tests: The first one is polymerase chain reaction testing and another one is antigen testing. Governments often make important decisions regarding testing protocols, procedures, and accessibility.

4.2.3. Theme 3: Public Response to Government Measures for COVID-19

Resources influence the ability of governments to provide assistance as well as their responsiveness to public needs and the public's trust, compliance, and overall response to the measures implemented. Effective communication is responsible for dispelling misinformation and building trust. This theme focuses on the impact of the following two topics.

4.2.3.1. COVID-19 Testing and Public Response.

The keywords "provision," "assistance," "response," "covid," and "policy" are interrelated in understanding the public response to government actions regarding COVID-19. Resources influence the ability of governments to provide assistance as well as their responsiveness to public needs and the public's trust, compliance, and overall response to the measures implemented. Effective communication is responsible for dispelling misinformation, building trust, monitoring mortality and infection rates, and reporting results.

4.2.3.2. Issues, Changes, and Guidelines on Public Health.

Social change as well as related issues may arise in the implementation of the new guidelines, based on our proposed study.

4.2.4. Theme 4: Social Influence

The social influence of the COVID-19 pandemic is enormous. It has perhaps changed daily life, the economy, and education widely. However, leadership, social, and public media have had a critical role in sharing pandemic-related information and ensuring public health. This theme focuses on the impact of the following four topics.

4.2.4.1. A New Lifestyle with the Pandemic.

The impact of the COVID-19 pandemic has led to lifestyle changes. Wearing surgical masks, washing hands, and other options to reduce the spread of the virus are some occasions that presented lifestyle behavioral changes.

4.2.4.2. Economic Changes.

"Economic" represents the economy of the country, and the keywords "business" and "education" represent important sectors of the economy. This highlights how governments implement restrictions and lockdown measures directly affecting businesses, such as supply chain disruptions and economic downturns. Elsewhere, businesses have adapted to the new normal, such as transitioning to e-commerce, contactless delivery, and virtual services. This all effected to drive economic growth.

4.2.4.3. Leadership and Media Influence for Increasing Public Health.

The "News" keyword references public media activities. Social media and public media have become critical features for sharing information and engaging in the decisions made by respective governments regarding pandemic-related measures and public health.

4.2.4.4. Effect of Education Section.

Although learning and teaching activities do not necessarily take place in the physical environment, continuing educational activities in a virtual environment is a positive change.

4.2.5. Theme 5: Vaccination

This discussion highlighted the importance of fair access to COVID-19 vaccines and the challenges faced in global vaccination efforts. It emphasizes medical professionals' and political leaders' partnerships to navigate medical guidance and government challenges. This theme focuses on the impact of the following two topics.

4.2.5.1. Global Vaccine Demand, Failures, and Distribution Inequity.

There is an inequality in global vaccine demand and distribution. The focus is on the efforts made by governments worldwide and the challenges they face in tackling the COVID-19 pandemic through vaccination strategies.

 https://doi.org/10.1633/JISTaP.2025.13.1.2

This involves discussing failures and emphasizing the need for more effective approaches to ensure equitable access to vaccines globally.

4.2.5.2. Role of Experts and Politics in Vaccination.

This topic emphasizes balancing the challenges of medical guidance and government. The COVID-19 vaccination campaign highlights the importance of maintaining a strong partnership between medical experts and political leaders.

Through this analysis, policies and measures taken to control the spread of the epidemic and ensure public health protection have been discussed as possible factors for political change during the COVID-19 epidemic. Examples include lockdowns, travel restrictions, vaccinations, and health precautions. As well, the underlying themes and topics that have been discussed on Twitter about the decisions taken by governments and the reactions given by citizens when they are implemented, as well as the trends that have occurred in society due to the COVID-19 epidemic, are revealed.

4.3. PLSA Topic Analysis

Considering the topic of the PLAS model, the model we generated 25 topics, and Table 4 shows a list of topics along with their corresponding probabilities in order from highest to lowest. Each topic represents a specific theme or concept, and the probability indicates the likelihood of that topic being present in a given dataset or document. These topics do not give an exact meaning to the dataset. However, this is based solely on the given word probabilities.

Table 4. Relevant words from the PLSA

Probabilistic latent semantic analysis (PLSA)			
Topic	Probability	Topic	Probability
'regulation'	0.0386	'supplier'	0.0126
'stage'	0.0385	'conflict'	0.0113
'distancing'	0.0361	'politics'	0.0112
'launch'	0.0351	'covid'	0.0109
'midnight'	0.0341	'bill'	0.0100
'update'	0.0271	'find'	0.0100
'stay'	0.0236	'http'	0.0098
'hey'	0.0189	'lack'	0.0083
'documentation'	0.0177	'interest'	0.0081
'store'	0.0158	'thanks'	0.0074
'procurement'	0.0137	'news'	0.0076
'contract'	0.0132	'transparency'	0.0071
'report'	0.0129		

The most frequent topics are "stay," "covid," and "news," highlighting the section on health during the COVID-19 pandemic. "Transparency," "regulation," "procurement," "conflict," and "politics" are keywords that refer to the political content of this data set. Others do not convey a specific meaning relating to our objectives. Most of them are verbs that are related to health protection. This means a significant portion of the analyzed data includes health concerns.

4.4. Word Cloud Visualization

The word cloud library is used here to illustrate the data visualization of pre-processed text from the data set. Words are arranged in no particular order, and the library randomly assigns colors to each word and visualizes word frequency and importance within a given text. Large words appear more frequently in the dataset, while smaller ones appear less often.

An overview of the 15 topics that the study generated is represented in Fig. 9. The size of each word reflects its frequency or importance within the corresponding topic. For example, the prominent words might include "covid," "suggest," "say," "good," "think," "go," "country," "give," "trust," "due," and "change." The word cloud provides an overview but does not convey detailed insights.

According to Fig. 10, "government," "covid," "good," "people," "say," "public," "health," and "decision" words are typically the larger words and are prominently shown rather than, "make," "change," "go," "failure," "tackle," "look," "demand," "globally," and "vaccine" words are prominently shown. Word clouds may not provide detailed information on word relationships, and they may not provide deep meaning to the dataset.

4.5. Analyzing the Results of Topic Modeling

Topic model evaluation determines how well a topic model performs its intended function. It is important to determine how well it performs the task for which it was created. When the topic model is used for a quantifiable goal, like classification, it is relatively easy to determine how effective the model is. However, evaluation is more challenging if the model is applied to a more qualitative task, such as examining the semantic themes in an unstructured corpus. Here, we evaluate the topic models using the coherence and Kullback-Leibler divergence (KL)



Fig. 9. Word cloud visualization in latent Dirichlet allocation.

scores.

4.5.1. Coherence Score

By calculating the degree of semantic similarity between high scoring terms in the topic, Topic Coherence measures the score of a single topic (Kapadia, 2019). These metrics help distinguish themes that can be understood semantically from topics that are the result of statistical inference. Among the different coherence measures, this study is based on the *C_v* measure. This measure builds word content vectors based on their co-occurrences and then determines the score using cosine similarity and normalized pointwise mutual information. The range from 0 to 1 is the *C_v* coherence score. Fig. 11 shows a graphical representation of the measurement of coherence scores. Higher scores of coherences are always better. This study identified that the optimal number of topics is 15, and its coherence value is 0.37.

4.5.2. Kullback-Leibler Divergence Score

The KL score, often known as the KL score (Heimel, n.d.), measures how different two probability distributions



Fig. 10. Word cloud visualization of the dataset.

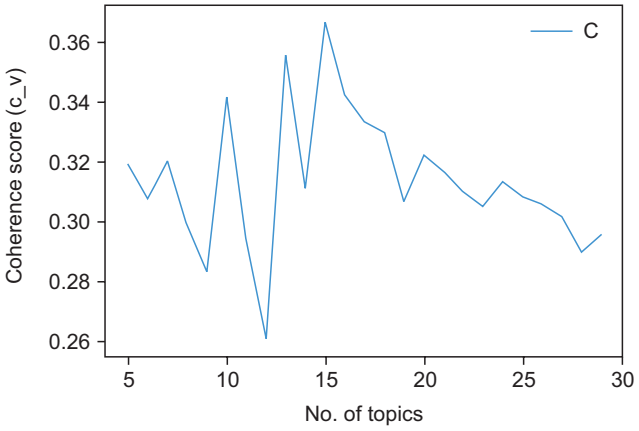


Fig. 11. Coherence score.

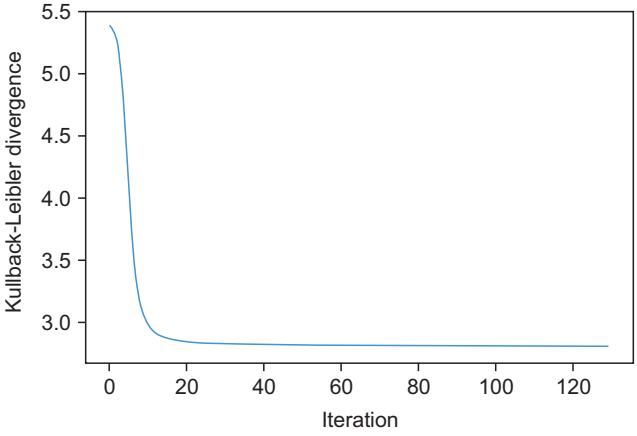


Fig. 12. Kullback-Leibler divergence score graph.

are from one another. KL divergence is a non-negative measure of the difference between two different probability distributions. When approximating a distribution, we can try KL divergence to reduce information loss (Avanesov & Kozlov, 2014).

The KL divergence is on a downward trend in Fig. 12, and the method is getting closer to its goal of minimizing the variance between the distributions. PLSA is hard to train, and this model is based on an extensive collection of documents. Therefore, we can use KL divergence to improve the time efficiency of the training algorithm. The curve's points represent the KL divergence values at particular iterations. Based on the lowest KL divergence attained, the optimal model can be chosen based on the KL divergence's stabilization. For instance, the KL divergence is 5.4 in the first iteration (y=0). Indicating that the distributions are becoming more similar or aligned, the KL divergence reduces throughout the course of repetitions.

Here, the coherence and KL scores are used to measure the "optimal number of topics" for the topic models. In this case, we selected 15 topics belonging to the values with the highest coherence score probability for the LDA, and the lowest score iteration was used, and 25 topics were selected in PLSA.

4.6. Discussion

When discussing the significance of our research findings in comparison to previous studies, the key point is that we were unable to find any prior research that specifically focused on the political changes during the COVID-19 pandemic using a topic modeling approach to extract relevant topics. However, when considering some relevant studies, they are limited to one country and use relatively small data sets. Also, those studies used some Google forms, interviews, discussions, weblogs, or poll data instead of social media data. This proposed approach filled this research gap. The proposed approach used global data targeting the whole world; the data set is relatively

large and successfully utilized both LDA and PLSA results for comparison. Moreover, Twitter data was more reliable for this kind of approach since people share their opinions on the platform freely and publicly. Therefore, when comparing the existing studies, the results of the proposed approach are significant and valuable for making more accurate decisions, ideas, and solutions for any relevant issues.

In this study, we extracted the main topics relevant to political changes during COVID-19 using a topic modeling approach. The topics that were identified as being related to political changes during the COVID-19 pandemic are explained in Table 3 and Table 4. As a result, 15 topics under five themes were identified using LDA (Table 3), and 25 topics were extracted through PLSA (Table 4). In order to identify topics, we used the most relevant keywords from each topic. For instance, consider topic no. 1, 'The travel mandates' in Table 3, where the keyword "travel" pertains to transportation. The keywords "government," "follow," "information," "mandate," and "guideline" act as a reminder of the rules and guidances to follow during the pandemic situation. On the other hand, the words "let" and "allow" imply restrictions on travelling. These kinds of suggestions are based on the topics and keywords in Table 3 and Table 4.

When comparing topic modeling approaches, LDA provided more accurate and coherent topics than PLSA. LDA is a valuable tool for comprehending the main discussions and trends surrounding political proceedings for health during the pandemic, as it provides a set of topics represented by their most probable words and their distribution across documents.

Important points excerpted from the Twitter data are that political authorities had to deal with many challenges and concerns brought on by the COVID-19 pandemic, and the focus was on the healthcare industry, which is grappling with the massive issues caused by the virus. The adoption of measures to prevent the spread of the virus and reduce its possible effects on individuals and communities depended heavily on political authorities. Addressing several issues, including virus detection and testing, manipulating vaccines, and healthcare infrastructure, has been given priority. In the discussion of the control of the spread of the virus, themes were cited along two lines: the effort to control the spread and the decision to control the spread of the virus. However, the adoption of decisions to prevent the spread of the virus and reduce its possible effects on individuals and communities depended heavily on political authorities.

Another concern is media influence and their leading actions to improve public health, affection of the economic and education sectors, lockdowns and travel restrictions, and how those issues have changed people's lifestyles, which frequently came into the discussion. Many users expressed doubts about the necessity and efficacy of these policies, as well as worried about how they might affect the economy and personal freedoms. Users discussed news, facts, and personal experiences about vaccination programs, marketing tactics, and vaccine reluctance. They also conferred with the experts to give their opinions about the significance of vaccination.

The online discussion regarding government decision-making regarding the process of vaccination highlighted the difficulty of managing a public health emergency and the significance of incorporating the general public into the decision-making process. Accordingly, those discussions were highlighted during the pandemic period and the proposed approach successfully extracted and finalized them under the different topics and themes.

4.6.1. Limitations of the Study

The study is constrained by Twitter data's inherent biases, which may not represent the broader public opinion due to demographic, geographic, and technical limitations. The findings reflect only the specific data collection period and may not capture the evolving political landscape. The dynamic nature of Twitter data also means that the findings are temporally limited and may not account for rapid changes in public discourse.

Language and cultural context are significant challenges, as the analysis might need to fully account for the nuances of multiple languages, dialects, or unique cultural expressions. Additionally, the research relies on the accuracy of NLP techniques, which can be impacted by the informal language, abbreviations, and slang commonly used on social media platforms. Furthermore, the study's scope is restricted to textual data, potentially overlooking multimodal content like images, symbols, and videos, which could provide a fuller understanding of political sentiment and discussions.

5. CONCLUSION

Identifying the main topics discussed related to political changes during the COVID-19 period is essential. Therefore, this study aimed to identify the most frequently discussed topics related to politics during the COVID-19 pandemic using the topic modeling approaches. This

study analyzed 10,658 political tweets during COVID-19 and political changes using Twitter between March 2020 and December 2021, applying LDA and PLSA topic modeling approaches. Both LDA and PLSA provided valuable insights into the latent topics present in the data, allowing us to understand the main discussions and trends related to political proceedings for health during the pandemic.

Through a comparative evaluation of the LDA and PLSA topic modeling techniques, we identified LDA as the superior method providing more accurate and coherent results. It helped to create a comprehensive explanation of the topics. It provided a set of topics represented by the most probabilistic words and the distribution of that topic across the documents. The adoption of measures to prevent the spread of the virus and reduce its possible effects on individuals and communities depended heavily on political authorities. Lockdowns, travel restrictions, governmental decision-making, vaccination and health measures, and lockdowns are some of the main topics of conversations on Twitter. Political authorities had to deal with several challenges and concerns brought on by the COVID-19 pandemic. Here, the focus was on the healthcare industry, which is grappling with the massive issues caused by the virus. The adoption of measures to prevent the spread of the virus and reduce its possible effects on individuals and communities depended heavily on political authorities.

Addressing several issues, including testing capacity, vaccine delivery, and healthcare infrastructure, has been given priority. Lockdowns and travel restrictions frequently came into the discussion. Political leaders' judgments on the duration, severity, and application of lockdowns were strongly criticized on the platform. The online debates surrounding travel bans, lockdowns, and government decision-making highlighted the difficulty in handling a public health emergency and the value of incorporating the general public into the decision-making process.

Twitter provided a platform for citizens to voice their opinions and hold political leaders accountable for their actions. Government decision-making processes also drew attention to this activity, with users on Twitter expressing both support and criticism for how leaders handled the pandemic. The knowledge from this research can be applied to similar pandemic situations to guide practical political behavior by the relevant political parties.

6. FUTURE WORK

In the future, we plan to introduce novel, innovative methods under the methodology section instead of using an existing methodology. This will help us to conduct an in-depth analysis of these topics without relying on the existing LDA and PLSA models. To further expand the data efficiently by using different data sources and different deep learning algorithms, we plan to apply a sentiment analysis approach to categorize the data.

The topic modeling research methodology contains some limitations, such as its lack of contextual understanding and the risk of overuse. It is crucial to delve into alternative strategies to address these challenges. As a solution for this, combining traditional topic modeling with artificial intelligence (AI)-based methods will help to improve interpretability and performance. For instance, integrating AI techniques can enhance topic modeling, revealing deeper and more subtle insights within textual data. Therefore, we plan to incorporate AI-based technologies with this topic modeling approach to improve the results in a future study. Furthermore, images, symbols, and videos, which could provide a fuller understanding of political sentiment, are planned to be integrated with textual data in future studies to gain a better understanding of opinions compared to the current study. These advancements will allow for a more detailed understanding of public sentiment, opinions, and reactions during pandemics, providing valuable insights for policymakers to make informed decisions in the future.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Alamoodi, A. H., Baker, M. R. Albahri, O. S., Zaidan, B. B., Zaidan, A. A., Wong, W. K., Salem, G., Albahri, A. S., Alonso, M. A., Jasim, A. N., & Baqer, M. J. (2022). Public sentiment analysis and topic modeling regarding COVID-19's three waves of total lockdown: A case study on movement control order in Malaysia. *KSII Transactions on Internet and Information Systems,* 16(7), 2169-2190. https://doi.org/10.3837/tiis.2022.07.003

Avanesov, V., & Kozlov, I. (2014, May 30). Faster PLSA. In N. Vassillieva, V. Ivanov, & D. Turdakov (Eds.), *Proceedings of the 10th Spring Researchers Colloquium on Databases and Information Systems* (pp. 2-9). CEUR.

Bermingham, A., & Smeaton, A. (2011, November 13). On using Twitter to monitor political sentiment and predict elec-

tion results. In B. Sivaji, & O. Manabu (Eds.), *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)* (pp. 2-10). Asian Federation of Natural Language Processing.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research,* 3, 993-1022. https://dl.acm.org/doi/10.5555/944919.944937

Bogdanowicz, A., & Guan, C. (2022). Dynamic topic modeling of Twitter data during the COVID-19 pandemic. *PLoS One,* 17(5), e0268669. https://doi.org/10.1371/journal.pone.0268669

Bose, R., Dey, R. K., Roy, S., & Sarddar, D. (2019). Analyzing political sentiment using Twitter data. In S. Satapathy, & A. Joshi (Eds.), *Information and Communication Technology for Intelligent Systems. Proceedings of ICTIS 2018* (Vol. 2, pp. 427-436). Springer.

Elghazaly, T., Mahmoud, A., & Hefny, H. A. (2016, March 22-23). Political sentiment analysis using Twitter data. In D. E. Boubiche, F. Hidoussi, L. Guezouli, A. Bounceur, & H. T. Cruz (Eds.), *Proceedings of the International Conference on Internet of things and Cloud Computing (ICC '16)* (Article No. 11). ACM.

Gandhi, A. (2021). *Topic modeling with LSA, PLSA, LDA & lda2Vec.* https://nanonets.com/blog/topic-modeling-with-lsa-plsa-lda-lda2vec/

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly,* 93(2), 332-359. https://doi.org/10.1177/1077699016639231

Heimel, G. (n.d.). *plsa.visualize module.* https://probabilistic-latent-semantic-analysis.readthedocs.io/en/latest/plsa.visualize.html

Howard, P. N., & Parks, M. R. (2012). Social media and political change: Capacity, constraint, and consequence. *Journal of Communication,* 62(2), 359-362. https://doi.org/10.1111/j.1460-2466.2012.01626.x

Howoldt, D., Kroll, H., Neuhäusler, P., & Feidenheimer, A. (2023). Understanding researchers' Twitter uptake, activity and popularity—An analysis of applied research in Germany. *Scientometrics,* 128(1), 325-344. https://doi.org/10.1007/s11192-022-04569-2

Kapadia, S. (2019). *Evaluate topic models: Latent dirichlet allocation (LDA)*. https://medium.com/towards-data-science/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Khotimah, D. A. K., & Sarno, R. (2019). Sentiment analysis of hotel aspect using probabilistic latent semantic analysis, word embedding and LSTM. *International Journal of Intelligent Engineering and Systems*, 12(4), 275-290.

Lee, E. J., Lee, H. Y., & Choi, S. (2020). Is the message the medium? How politicians' Twitter blunders affect perceived authenticity of Twitter communication. *Computers in Human Behavior,* 104, 106188. https://doi.org/10.1016/j.chb.2019.106188

Lim, W. M. (2021). Editorial: History, lessons, and ways forward from the COVID-19 pandemic. *International Journal of Quality and Innovation,* 5(2), 101-108. https://www.inderscience.com/info/dl.php?filename=2021/ijqi-7348.pdf

Lipscy, P. Y. (2020). COVID-19 and the politics of crisis. *International Organization,* 74(S1), E98-E127. https://doi.org/10.1017/S0020818320000375

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval,* 14(2), 178-203. https://doi.org/10.1007/s10791-010-9141-9

Mameli, M., Paolanti, M., Morbidoni, C., Frontoni, E., & Teti, A. (2022). Social media analytics system for action inspection on social networks. *Social Network Analysis and Mining,* 12(1), 33. https://doi.org/10.1007/s13278-021-00853-w

Manthiramoorthi, M. (n.d.). *Topic modelling techniques in NLP*. https://iq.opengenus.org/topic-modelling-techniques/

Meduru, M., Mahimkar, A., Subramanian, K., Padiya, P. Y., & Gunjgur, P. N. (2017). Opinion mining using Twitter feeds for political analysis. *International Journal of Computer,* 25(1), 116-123. https://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/929

Montiel, C. J., Uyheng, J., & Dela Paz, E. (2021). The language of pandemic leaderships: Mapping political rhetoric during the COVID-19 outbreak. *Political Psychology,* 42(5), 747-766. https://doi.org/10.1111/pops.12753

Olbrich, E., & Yamshchikov, I. (2018). *Statistical inference of political spaces*. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bbe1d647&appId=PPGMS

Palmer, C. (2021). *Political attitudes and behavior in a pandemic: Factors affecting compliance with COVID-19 policies* (master's thesis). University of Tennessee, Knoxville, TN.

Parack, S. (2021). *Introducing the new academic research product track*. https://devcommunity.x.com/t/introducing-the-new-academic-research-product-track/148632/1

Perera, S., & Karunanayaka, K. (2022, February 23-24). Sentiment analysis of social media data using fuzzy-rough set classifier for the prediction of the presidential election. *Proceedings of the 2nd International Conference on Advanced Research in Computing* (pp. 188-193). IEEE. https://doi.org/10.1109/ICARC54489.2022.9754173

Purushothaman, U., & Moolakkattu, J. S. (2021). The politics of the COVID-19 pandemic in India. *Social Sciences,* 10(10), 381. https://doi.org/10.3390/socsci10100381

Rathnayaka, R. M. D. G., & Rupasingha, R. A. H. M. (2023). Management lessons from impact of Covid-19 on political changes: Sentiment analysis using Twitter data. *IUP Journal of Management Research,* 22(2), 5-27. https://iupindia.in/ViewArticleDetails.asp?ArticleID=6784

Rieger, M. O., & Wang, M. (2022). Trust in government actions during the COVID-19 crisis. *Social Indicators Research,* 159(3), 967-989. https://doi.org/10.1007/s11205-021-02772-x

Sievert, C., & Shirley, K. E. (2014, June 27). LDAvis: A method for visualizing and interpreting topics. In J. Chuang, S. Green, M. Hearst, J. Heer, & P. Koehn (Eds.), *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63-70). Association for Computational Linguistics.

World Health Organization. (n.d.). *Coronavirus disease (COVID-19).* https://www.who.int/emergencies/diseases/novel-coronavirus-2019

Wright, L., Burton, A., McKinlay, A., Steptoe, A., & Fancourt, D. (2022). Public opinion about the UK government during COVID-19 and implications for public health: A topic modeling analysis of open-ended survey response data. *PLoS One*, 17(4), e0264134. https://doi.org/10.1371/journal.pone.0264134

Yano, T., Cohen, W. W., & Smith, N. A. (2009, May 31-June 5). Predicting response to political blog posts with topic models. In M. Ostendorf, M. Collins, S. Narayanan, D. W. Oard, & L. Vanderwende (Eds.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 477-485). Association for Computational Linguistics.

JISTaP https://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 13(1): 36-50, 2025

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2025.13.1.3

Library Space and Resource Usage Time: Reflections from Digital Footprints During Pre and Post-COVID-19

Wannisa Matcha
Department of Computer and Informatics for Management, Faculty of Communication Sciences, Prince of Songkla University, Pattani, Thailand
E-mail: wannisa.ma@psu.ac.th

Nawapon Kaewsuwan*
Department of Information Management, Faculty of Humanities and Social Sciences, and Research Center for Educational Innovations and Teaching and Learning Excellence, Prince of Songkla University, Pattani, Thailand
E-mail: nawapon.k@psu.ac.th

Komgrit Rumdon
John F. Kennedy Library, Office of Academic Resources, Prince of Songkla University, Pattani, Thailand
E-mail: komgrit.r@psu.ac.th

ABSTRACT

This paper aims to investigate library user behaviors by comparing them pre and post-COVID-19. Two aspects of users' behaviors were examined, including using physical learning spaces, and online resource utilization. This exploration used secondary data automatically recorded by ALIST OPAC systems (Prince of Songkla University Pattani Campus, Pattani, Thailand) during 2018 and 2022. Descriptive statistics were used to examine physical space usage. The machine learning algorithms name process mining technique was used to investigate the changes in online behaviors, by formulating the sequence of activities based on the digital footprints recorded when users interacted with the systems. This technique exemplified the sequence and timing of activities. The study results revealed an increased use of limited co-learning spaces by library users after COVID-19. The activity of borrowing and returning library resources showed that fewer reminder e-mails on the due date were sent. The process of mining exemplified the sequence and timing of activities. It indicated that after COVID-19, library users hold borrowed items for a shorter time than before the pandemic. These findings suggest that the digital footprints unveiled the changes in library users' behaviors. That is, the library users return borrowed items faster than in pre-COVID-19 circumstances. A decrease in reminder e-mails was clearly visible to support such a finding. Therefore, it is suggested that library managements should consider a faster operation and well-resourced management to adapt to the changes. Managing resources needs to become faster as the library users showed a potential for faster return of borrowed resources.

Keywords: library spaces, user behaviors, digital footprints, data mining, process mining, COVID-19

Received: February 7, 2024 **Revised:** September 12, 2024
Accepted: October 15, 2024 **Published:** March 30, 2025

***Corresponding Author:** Nawapon Kaewsuwan
https://orcid.org/0000-0002-0454-3972
E-mail: nawapon.k@psu.ac.th

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. Open Access articles are automatically archived in the Korea Institute of Science and Technology Information (KISTI)'s Open Access repository (AccessON). All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/). Under this license, the authors retain full ownership of their work, while permitting anyone to use, distribute, and reproduce the content in any medium, as long as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© 2025 [Wannisa Matcha, Nawapon Kaewsuwan, Komgrit Rumdon] This article is distributed under the Creative Commons Attribution License (CC BY), allowing unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. It is published by the Korea Institute of Science and Technology Information (KISTI).

1. INTRODUCTION

The global pandemic of coronavirus, also known as COVID-19, has had great effects in all aspects (Al-maaitah et al., 2021). That is, the dangerous spread of the virus resulted in the restriction of physical social interaction (De Groote & Scoulas, 2021; Marome & Shaw, 2021). Hence, all activities involving human interaction had to be paused. This resulted in the exponential use of technology. Consequently, several new technologies and applications were developed to assist humans in continuing their work (Ćirić & Ćirić, 2021; Yu et al., 2023). People were pushed to adapt to the new way of life. Many become much more comfortable communicating via video conferencing, reading, and working through digital devices. Many resources became available in the digital format. Yet, some were keen on physical interaction and aimed to make up for the time they lost during the shutdown (Wassler & Fan, 2021). The latter is referred to as the 'revenge user's effect.' This term is frequently used in several fields. For instance, revenge tourists were used to explain a surge of tourists after COVID-19 in several countries such as Indonesia (Abdullah, 2021) and India (Joshi & Sadhale, 2022). These surges were influenced by the need for physical travel and to make up for the time people had to stay at home during the pandemic (Wassler & Fan, 2021). This similar situation was expected in various areas such as business, marketing, and education, as well as public spaces and co-working spaces like libraries.

The disruption of COVID-19 changes users' behaviors (Al-maaitah et al., 2021). In a Thai context, the implementation of movement restrictions, social distancing, and promotion of online working has a significant contribution and is embedded in the new normal routine (Mahapornprajak, 2021; Marome & Shaw, 2021). According to a research survey conducted by Mahidol University (2020), after COVID-19 the Thai population was strongly aware of and positively responded to the campaign to prevent and control the virus spread. Thai citizens paid close attention to social distancing regulations, as reflected by their decision to stay at home or in limited closed spaces, such as a small meeting room, well air-circulated spaces, and limited numbers of entry spaces considered safe from virus spreading. Their study revealed that the Thai population tried to limit time spent in public and crowded spaces such as parks, restaurants, and working spaces as well as libraries (Marome & Shaw, 2021). The studies presented thus far reflect the changes in users' perceptions and behaviors in general. However, much is unknown regarding the behaviors of users when using the library. It is hypothesized that the behaviors of library users changed after COVID-19. However, to what extent does this change affect the operation of the library? This question remains unanswered.

Before COVID-19, many libraries in Thailand allowed physical borrowing of books at the library to verify the identity of the user, checking out books, and returning of borrowed books. During COVID-19, this operation was replaced with online booking and delivery systems. Nevertheless, there is a lack of research specifically focused on the changes in library users' behaviors after COVID-19, especially for Thai citizens.

The library is a large source of knowledge and learning space (Unchitti, 2005). The library is managed by many organizations, such as information centers, specialized libraries, national libraries, and public libraries, as well as libraries in educational institutions, including school libraries and university libraries. The library plays an essential role in knowledge acquisition and development for all education levels. Not only does the library provide information, but it also offers other support, including borrowing and returning resources, information management tools, learning and co-working spaces, information searching and management consultations, and many more activities (Kamolpechara & Soopunyo, 2020; Unchitti, 2005).

The growth of technology and the hit of COVID-19 have impacted library usage (Ćirić & Ćirić, 2021; De Groote & Scoulas, 2021). Libraries were forced to close during COVID-19 (Salvesen & Berg, 2021). Several libraries developed online systems to keep their services running and facilitate access to learning resources (De Groote & Scoulas, 2021; Yu et al., 2023). Some developed book and library resource delivery systems to support distance learning (Yu et al., 2023). Not only do these changes affect users in terms of services, but COVID-19 also might impact users' behaviors. For instance, users might come to prefer reading digital files rather than physical books (Millar & Schrier, 2015). They might also spend more time reading (Alomari et al., 2023). Yet much is unknown regarding user behaviors (Alomari et al., 2023). Hence, it is important to understand the changes in users in order to provide effective services to them (Sangchantr, 2022), especially for university libraries, which offer their services to different groups of users such as higher education students, researchers, academicians, and the public. Studying the university library users' behaviors recorded by the system is essential for the library to modify its op-

eration to suit the needs and behaviors of its users (De Groote & Scoulas, 2021). This could also provide suggestions on how to allocate resources efficiently. Moreover, user behavior analysis could also lead to the effective use of available resources and reduce unused resources in the long run.

Generally, libraries record a high volume of data (Horstmann & Brase, 2016). Yet this data is still suboptimally used. In fact, the recorded data contain the digital footprint of users when interacting with the system. This is referred to as 'trace data.' Several fields of study utilize this data to enhance the process. However, there is a lack of study focusing on trace data obtained from library management systems, much less examination of the changes in user behaviors. By in-depth analysis of the trace data, it is expected that some valuable insights into library users' behaviors after COVID-19 could be observed. These insights can be utilized to enhance the service quality of university libraries or information service institutions in the post-COVID-19 era, particularly in the process of library operation. Hence, this research aims to bridge the gap by comparing changes in terms of physical space usage and online operation during the pre and post-COVID-19 period. Examining such trace data requires an advanced machine learning (ML) algorithm due to the large volume of data. Therfore, process mining, which was built on ML algorithms to discover the process, was used. Previous study showed that the process mining algorithm enables the best practice in library operation (Dzihni et al., 2019; Kouzari & Stamelos, 2018).

Moreover, the research outcomes and recommendations can be applied in policy and strategic planning for information services, budget allocation, service organization, and online or electronic information dissemination formats that align with the changing behaviors of individuals. These findings can inform the design and provisioning of information services and learning spaces that are suitable and user-friendly. Additionally, they can assist in sharing best practices for data management to enhance future library management systems.

2. OBJECTIVES

To analyze and compare library users' behaviors before and after COVID-19 by examining physical space usage and online resource usage time.

3. LITERATURE REVIEW

3.1. Library Transformation: Traditional Libraries Heading into the Artificial Intelligence Era

Universities often organize their libraries to offer open information resources for users. Traditional libraries are physical institutions that serve as a repository of information resources. They are spaces where users can access print resources such as books, journals, newspapers, magazines, and other materials for educational, recreational, research, and other purposes. These libraries often have several types of spaces: quiet reading areas, individual learning spaces, and collaboration and discussion spaces. There are librarians who assist users in retrieving and utilizing the available information resources and systems (Ashikuzzaman, 2023).

Recent developments in technology have had a high impact on the library, as many users are promptly adapting to digital libraries. This is because these resources are flexibly accessible via the Internet. They are also known as online resources. User behaviors have changed, resulting in an increased number of digital reading materials and digital media. However, university libraries are still considered vital learning spaces. According to Asif and Singh (2020), traditional libraries have been turned into digital libraries due to technical developments. Libraries have a wide variety of e-resources, e-services, and other functions, as well as distributing information in e-formats to meet the information needs of modern library users.

Over the past decade, libraries have developed their activities and affordance with the use of technology. For example, augmented reality and virtual reality, emerging technologies that utilize virtual objects to enhance reality, have been used to improve the library user experience and engagement. Several types of software, applications, and systems have been developed to support the operation of the library and user education (Oyelude, 2017; Yu et al., 2023). Recently emerged generative artificial intelligence tools, such as ChatGPT, have also been integrated into the library system. According to Ehrenpreis and DeLooper (2022), using an AI Chatbot on library websites to answer questions helps retrieve or recommend information sources like a librarian.

3.2. Reading Behaviors

Various unexpected changes have occurred due to the great impact of the COVID-19 pandemic. These changes have affected many sectors, mostly in social, economic, and education domains. Consequently, people were

 https://doi.org/10.1633/JISTaP.2025.13.1.3

strongly urged to stay home, and even work and study are transformed into online modes (MacKenzie, 2020). Libraries had to be physically closed. To some extent, their services were provided by online systems. At the same time, digital content became the sole option for these libraries' users (Ćirić & Ćirić, 2021). As a result, user behavior has changed with technology that supports easier access to information, and users can access it anywhere, anytime, through the Internet. As a consequence, the rapid ongoing transition of reading from print to screen is clearly visible. This creates challenges to the use of physical books. At the same time, it creates the opportunity for digital books and reading materials due to an increasing number of digital reading devices, such as computers and laptops, tablets, smartphones, and other forms (Horstmann & Brase, 2016).

The paradigm of reading, in particular for users, is increasingly screen-based rather than paperbound. As Internet use has developed increasingly, more reading information resources have transformed into digital format, gaining more readers through this platform. In contrast, paper-based reading also has its particular benefits, such as time efficiency (Delgado et al., 2018), eye-friendliness (Tosun, 2014), comprehensibility (Duran, 2013; Jabr, 2013), and simplicity (Jabr, 2013). Regardless of the fact that some people refuse to use paper due to the increasing campaign of wasting less paper (Parlindungan et al., 2021), many people still find paper their preferred reading material. Additionally, some research has argued that digital environments may not always be best suited to fostering deep comprehension and learning. For example, Mangen et al. (2013) found that students who read texts in a printed format scored significantly better on the reading comprehension test than students who read the texts digitally. Similarly, Delgado et al. (2018) suggested that providing students with printed texts, despite the appeal of computerized study environments, might be an effective way to improve comprehension outcomes. Therefore, the role of libraries, which contain a large number of paperbound materials, are essential for education.

However, given the unexpected attack of COVID-19, users' reading behaviors and library usages are expected to change. For instance, Ćirić and Ćirić (2021) found that the average reading time of participants from Serbia increased by approximately 130% during the lockdown. Baharuddin and Mohamad Rosman (2020) further found that information quality and responsiveness are two important factors that predicted potential library e-services utilization during COVID-19. However, these studies focused on the reading behaviors of digital text by using data gathered from a survey. There is a lack of research utilizing recorded trace data to examine the factors or behaviors of library users during the post-COVID-19 era. Hence, this present research aims to investigate how the behaviors of library users have changed from pre- to post-COVID-19.

3.3. Trace Data Analysis

Interaction with the available information systems offered by the library generates a large volume of data. This data is often used to understand the behavior of users in various fields. For example, trace data is recognized as a highly insightful source of information in education (Gašević et al., 2015). It has been reported as a more reliable source as compared to the use of self-report instruments such as questionnaires, interviews, or think-aloud protocols (Srivastava et al., 2022; Zhou & Winne, 2012). Trace data has also been used in business to understand customer behaviors (Kröckel & Bodendorf, 2012). Similarly, it was used to explore the performance conformation of the hospital process (Gao et al., 2020). As presented thus far, trace data records insightful information about users' interactions that could be used to strengthen the understanding of user interactions with the system.

To understand the library user and analyze activity relevant to the library, trace data could be used. However, trace data records a large amount of data; a simple count might not offer actionable insights to improve an existing activity and its relevant activities. ML as well as data mining algorithms are often used to unveil the insight hidden within the trace data. ML and data mining algorithms are mathematics-based algorithms that can be grouped under the umbrella of AI. It can be programmed to order the computer to automatically retrieve, recognize, identify, classify, or predict some information from the given data set.

One of the ML algorithms used to mine trace data is process mining. Process mining enables automatically creating a process model from event logs without prior knowedge of the process and analyzing the performance of processes, such as identifying bottlenecks, delays, or resource inefficiencies. Originally, this technique was used to analyze and improve business processes by extracting information from event logs and trace data generated by information systems (Janssenswillen et al., 2019). Recent literature also applies such methods to understand the activity flow of the library process. For instance, Kouzari and Stamelos (2018) applied process mining to examine the differences in operational practices across libraries using

the same integrated library system. They found that even though two organisation use the same system, there were differences in terms of practices, activities, sequences, and approaches followed in daily tasks. Such findings could be used to improve existing practices. Similarly, Dzihni et al. (2019) used process mining to examine the bottleneck in the helpdesk process of a university's academic information system. They were able to indentify two activities in the helpdesk process that required improvement, which were activities during input, and seeing the ticket or request.

Despite the usefulness of process mining, to the best of our knowledge there are a very small numbers of research studies using process mining to examine the processes and activities of the library, especially library users' behaviors that might have changed due to COVID-19. Therefore, this research aims to explore the changes in library users' behaviors while using the library resources by comparing pre and post-COVID-19. The behaviors of library users were automatically recorded when they interacted with the system. Therefore, two dimensions of library usage were explored, including library physical access and facilities reservation, and online behaviors.

4. RESEARCH METHODOLOGY

4.1. Library Context

This study focused on library user visitation and utilization of the resources provided by the university library. The secondary data from the system's recorded interactions were used. The library focused on in this study is one of Thailand's southernmost university libraries. The university has several campuses with their own libraries. The library offers services to several types of users,

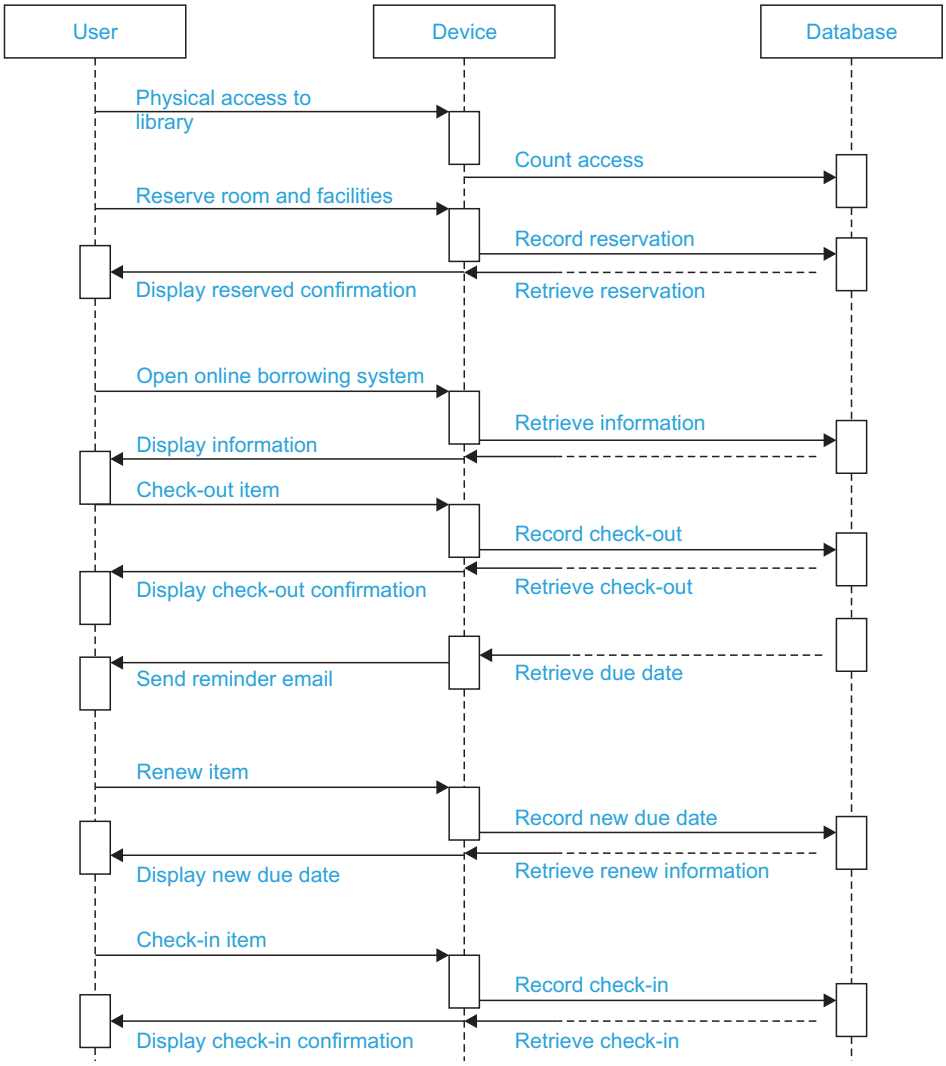


Fig. 1. Sequence diagram illustrating sequences and supported systems relevant to this study (own data processing).

 https://doi.org/10.1633/JISTaP.2025.13.1.3

including academics, researchers, graduates, undergraduates, university employees, and external university users, including adults and children. There are eight main faculties offered on the campus. Library users can physically visit the library during working hours from 8:30 to 20:30 daily. During examination weeks the library offers 24-hour access. When users pass through the entry gate, the automated system will read the ID and start counting the entry.

The university's library offers a variety of resources. This includes paperbound and digital-based reading materials, audio and video materials, computers, working spaces, meeting rooms, discussion rooms, and electronic devices (Prince of Songkla University Pattani Campus, n.d.). Some of the resources are made available via online systems. Some resources are physically available at the library. This study focused on items that were made available to physically borrow at the university, such as paper-based books, meeting rooms, and facilities.

However, borrowing items from the library must be done through an electronic system. In general, graduates and external users can borrow these resources for seven days. They can renew the borrowed resources by themselves via the online system. All borrowed items need to be physically returned to the library by a specific time or else the user will be fined. Undergraduate students, university employees, and academicians are allowed to borrow items for 30 days and renew them through the online system. In 2018, the renewing function was limited to one time per item. However, in 2022 multiple renewals were allowed. An e-mail reminder will be automatically sent to users if they approach the due date.

4.2. Data Analysis

A large volume of data is recorded via the library system. The collected data in 2018 and 2022 were used to understand the behaviors of library users during pre- and post-COVID-19. The data were integrated from several resources to analyze the user's digital footprint left behind when interacting with the library systems. Fig. 1 presents the relevant sequence of the use of library resources.

Physical entry to the library was recorded by using the university's self-developed system, called the ASSO system (Prince of Songkla University Pattani Campus, Pattani, Thailand). Only anonymized IDs and the number of times users performed specific actions, such as accessing the library or checking out items, were used in this study. The reservations for meeting rooms, collaborative working spaces, computers, and other facilities were recorded via the room booking system. Only anonymized IDs and a number of reservation counts were used. Borrowing and returning the items are operated via the university's self-developed system, called ALIST OPAC (Prince of Songkla University Pattani Campus, Pattani, Thailand). Log data recording the user's interaction with the system was used. Based on the exploration, the main learning events recorded by the system included check-in items, check-out items, reminder e-mails, and renewal items. Also, some minor events were recorded, such as inserting, updating, deleting the remark, holding, canceling the hold item, and reporting the lost item. A brief description and its corresponding event are presented in Table 1.

Table 1. Event and description of events recorded by the online library system (own data description)

No.	Activity	Description
1	Check-out item	A user borrowed and checked out the item from the library
2	Check-in item	A user checked in and returned the book to the library
3	Reminder e-mail	The library system sent a reminder e-mail about the due date of the borrowed item to the corresponding user
4	Renew item	A user renewed the borrowed item and extended the due date
5	Update due datetime	The system updated the due date manually
6	Insert block message	A user inserted a remark note
7	Delete block message	A user deleted a remark note
8	Update check-outs note	A user updated the check-out remark note
9	Hold item	A user queued to borrow a specific item
10	Cancel hold item	A user canceled the queue
11	Lost item	A user reported a lost item

To prevent any identification of library users, only the anonymized ID, corresponding department, type of user, events, borrowed items, and timestamp were used to understand the behaviors of the users.

The university library recorded a huge amount of data; hence, to understand the users' behavior extensively, the ML algorithm was used to analyze the data. In this study, process mining was used to formulate the sequence of activities of library users. Before applying process mining, the trace and its sequences were explored by the bupaverse R package (Janssenswillen et al., 2019). The trace explorer function enables us to visualize sequences within the coverage ranges, most frequently observed sequences, and infrequent sequences. Then, the bupar R library (Janssenswillen et al., 2019), one of the most frequently used R libraries, was used to explore the activities and sequence of activities.

The sequence of activities was illustrated by using a directly-follows graph (DFG). DFG is the simplest and most commonly used method to formulate the process based on the occurring event (Van Der Aalst, 2012). It simply requires an activity instance, sometimes referred to as case, activity, and timestamp. Some other variables are optional when computing the diagram. To address the research aim, the sequence of activities performed by library users was formulated for each iteration of the borrowed item by each user. Hence, the case variable referred to any related events performed by a user from borrowing until returning the specific item. The activity referred to the event recorded in the system as presented in Table 1. The timestamp refers to the specific time when the event occurred.

As a result of using DFG via bupaR R library, the sequence of activities was illustrated. It is composed of nodes, numbers, and flows (arrows). Each node represents each activity, as presented in Table 1. The flows of activities were illustrated by using the arrows to represent the direction from the source activity to the next activity. Hence, the behaviors derived from user interactions prior to COVID-19 and after COVID-19 can be examined.

5. RESULTS

5.1. Accessing and Facilities Reservation

Exploration into library usage before and after COVID-19 revealed a decreased number of library users. Prior to the global pandemic, 9,674 library users physically visited the library in 2018. The majority of them were undergraduates (5,314, 54.93%), academics and researchers (3,394, 35.08%), external university users (397, 4.10%), graduates (328, 3.39%), and university employees (242, 2.50%). A total number of 192,920 counts were recorded for physical entry to the library. The average number of visits was 11.33 times. However, in 2022 after the global pandemic, the number of library users who physically visited the library decreased to 7,511. Similarly to the pre-COVID-19 period, most of them were undergraduates (4,442, 59.14%), academics and researchers (2,164, 28.81%), external university users (654, 8.71%), graduates (189, 2.52%), and university employees (62, 0.82%). The visitation counts came to 94,768 times. The average number of visits was 7.57 times. The distribution of visitation counts showed similar patterns, as presented in Fig. 2.

Even though the number of visits decreased, the number of meeting rooms and co-working space reservations increased. That is, in 2018, 675 reservations were recorded

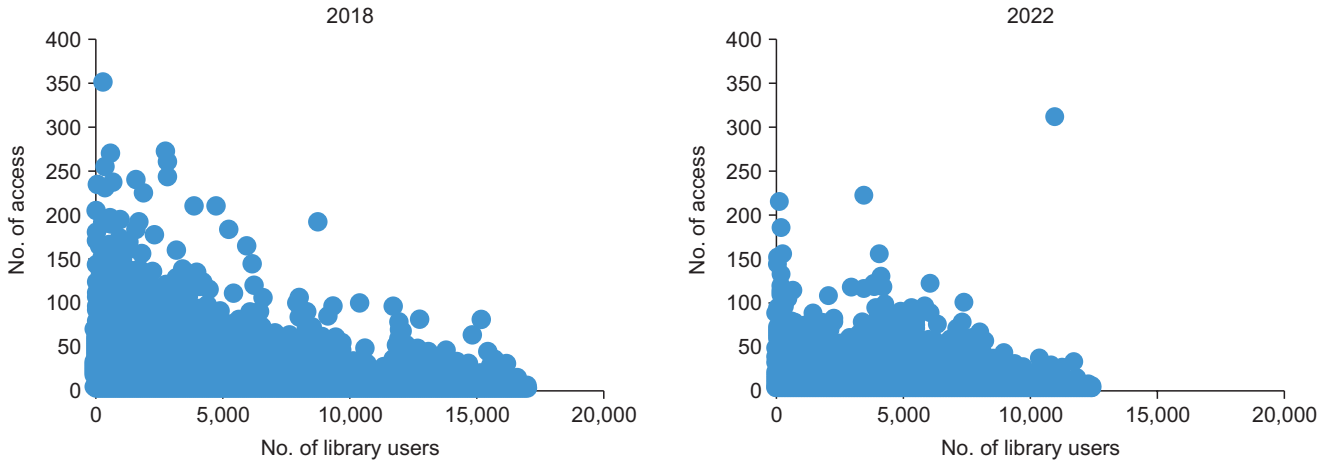


Fig. 2. Number of times library users physically visited the library (own data processing).

 https://doi.org/10.1633/JISTaP.2025.13.1.3

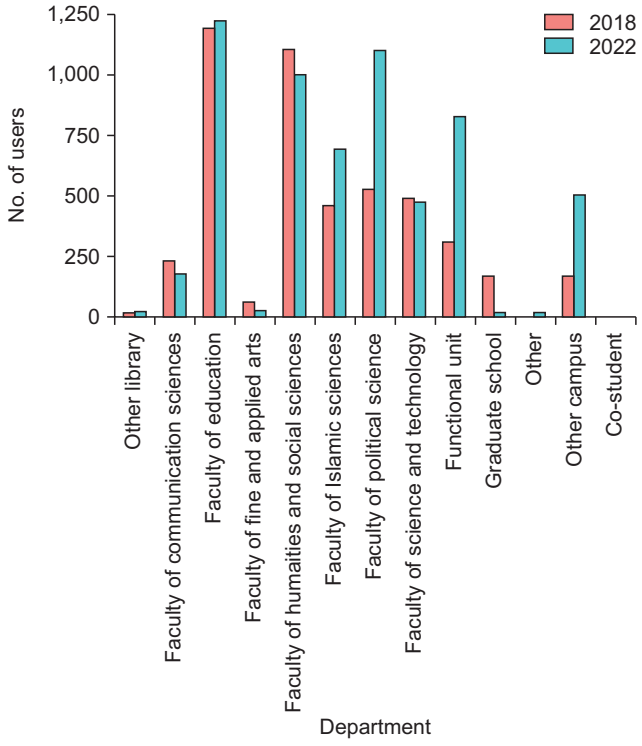


Fig. 3. Number of library users (own data processing).

in the database. In 2022, the number of reservations increased to 1,552.

When looking into online behaviors (refer to Fig. 3), it was revealed that most library users were undergraduate students enrolled in the faculty of education, humanities and social sciences, and political sciences. In 2022, there was a very large increase in library users from the faculty of political sciences, the faculty of Islamic sciences, and those from functional units as well as from other campuses.

5.2. User Behaviors When Interacting with the Online Library System

Fig. 4 presents the sixteenth most frequently occurring sequences of interaction exhibited by library users. These sequences covered 13% of all sequences. As shown in Fig. 4, the most frequently observed sequence was checking out items from the library and returning the items. Other patterns that frequently occurred were checking out items from the library, followed by the reminder e-mail regarding the due date, and then returning the item to the library after receiving the reminder e-mail. When looking at the whole diagram, it is clear that e-mail reminders were the most frequently occurring event when interacting with

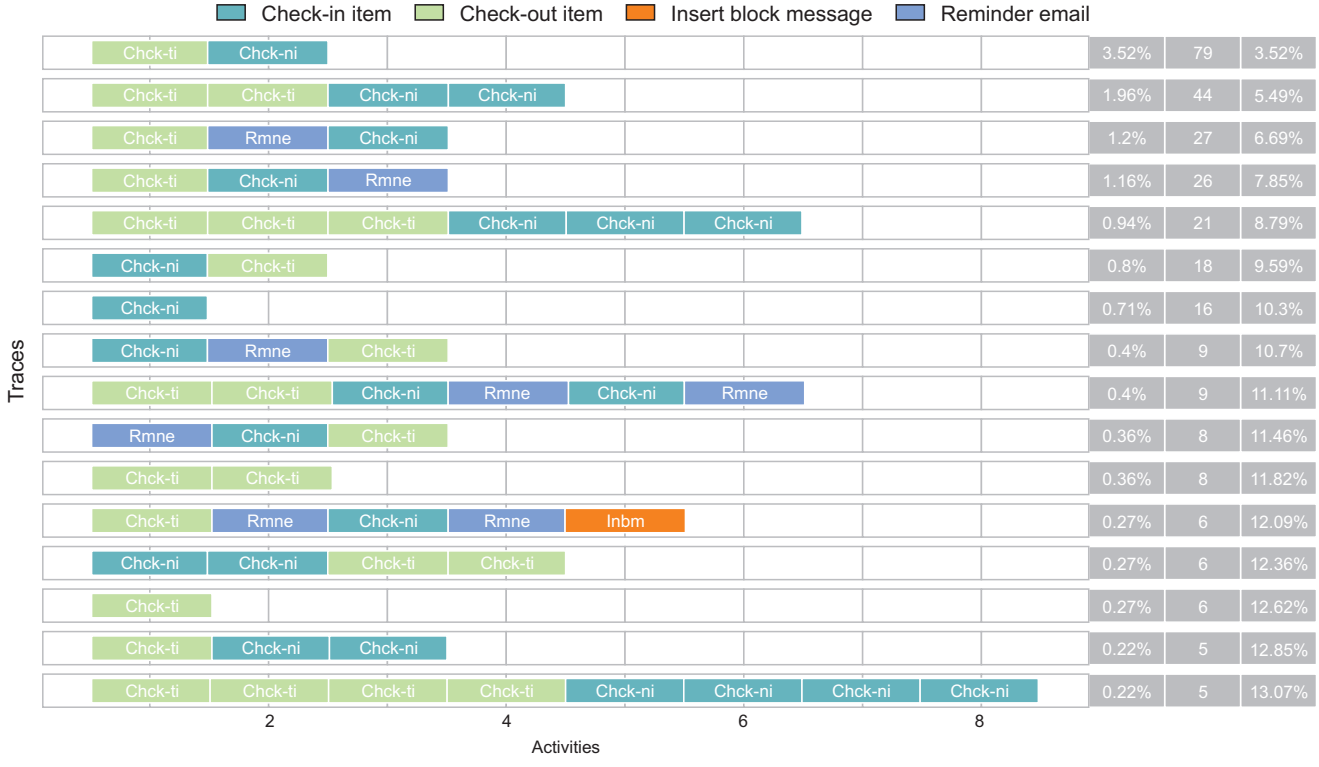


Fig. 4. Sixteenth most frequently occurring activity in 2018 (own data processing).

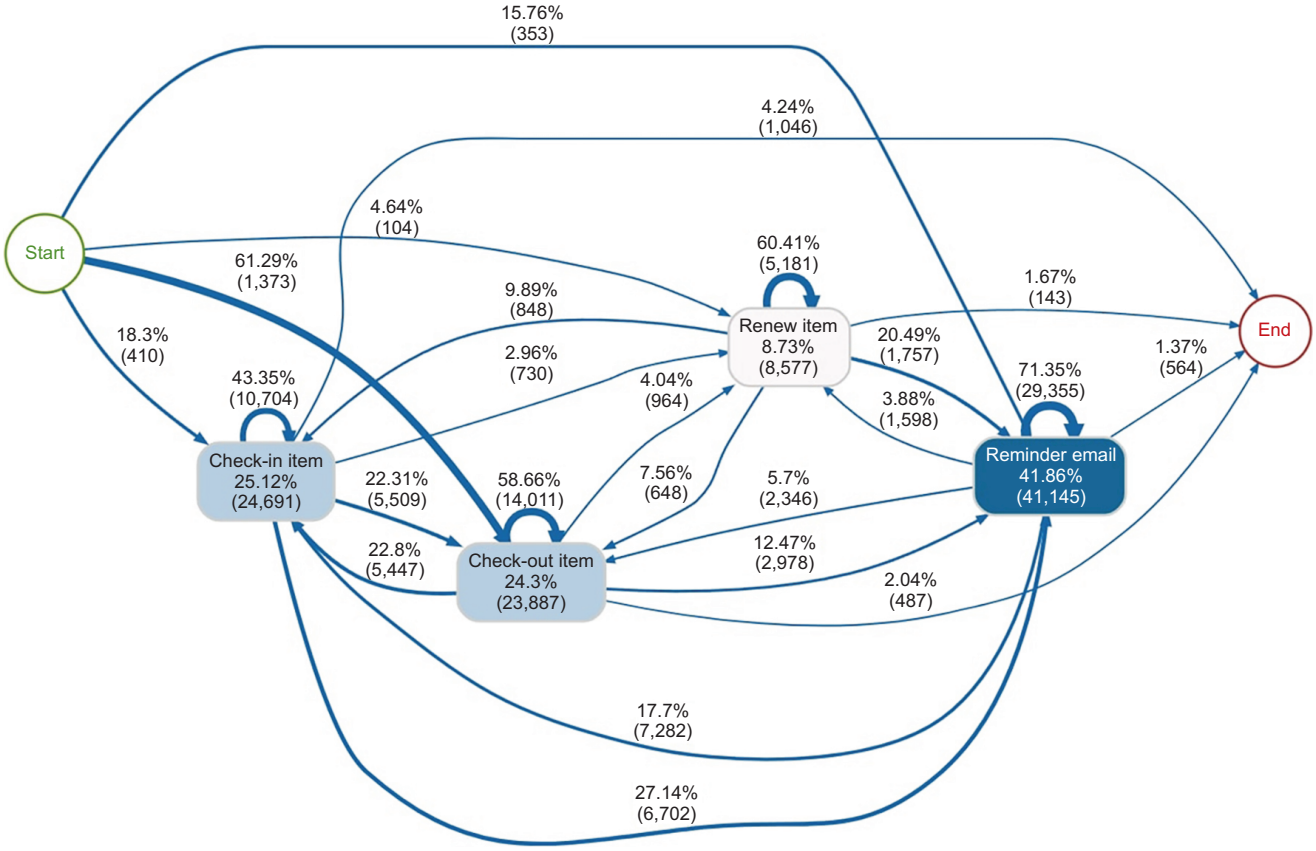


Fig. 5. Overall activity sequence of user interaction with the online library system in 2018 (own data processing).

the ALIST system, as presented in Fig. 5. The node in Fig. 5 represents the type of interaction with the ALIST system. The darker shade of node color represents the higher records of such interactions. The percentage in the node indicates the proportion of interaction against all recorded interactions. The numbers located inside the node show the actual number of recorded interactions. The arrow indicates the flow of the activity from the one interaction and its following interaction. The stronger and thicker the line signifies the higher frequency of occurrence and vice versa.

In 2018 (refer to Fig. 5), most library users began their activity by borrowing items (start → check-out item) from the library. Note the beginning of interactions by returning the borrowed items (start → check-in item) to the library, as this action might be a result of continued activity from the previous year. The proportion of interactions between checking-in (check-in item → check-out item) and checking-out items (check-out item → check-in item) from the library is almost equal (22.31% and 22.80%, respectively). This indicates that the user will normally borrow the items from the library and return them after some time. However, the strongest, darkest shade of nodes is the reminder e-mail (accounting for 41.86% of all recorded interactions), showing a strong direction to check-in (reminder e-mail → check-in item) and check-out (reminder e-mail → check-out item) of items from the library. This means that most of the time, library users receive a reminder e-mail regarding their borrowing, returning, and due date of the items.

Similar to 2018, in 2022 the most frequently occurring activities were checking out items from the library and returning the items to the library. This sequence accounted for 9.61% of all detected sequences. Also, library users often borrowed multiple items at a time, as indicated by consecutive check-outs followed by separate check-ins. The main difference observed was that in 2022, the number of e-mail reminders to return books was much lower than in 2018. The interaction of renewed items was also observed in the 16 most frequently observed sequences (Fig. 6).

Fig. 7 presents the overall sequence of activities based

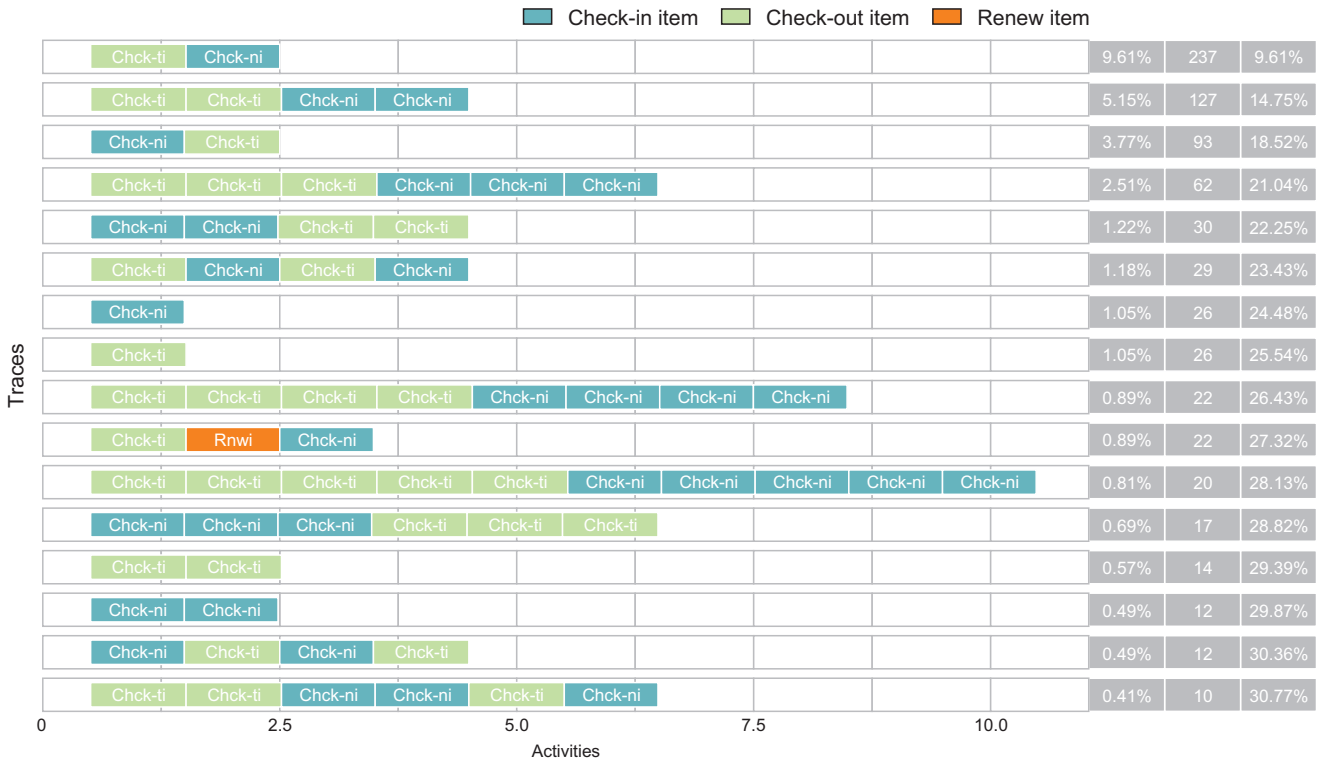


Fig. 6. Sixteenth most frequently occurring activity in 2022, covering 31% of sequences (own data processing).

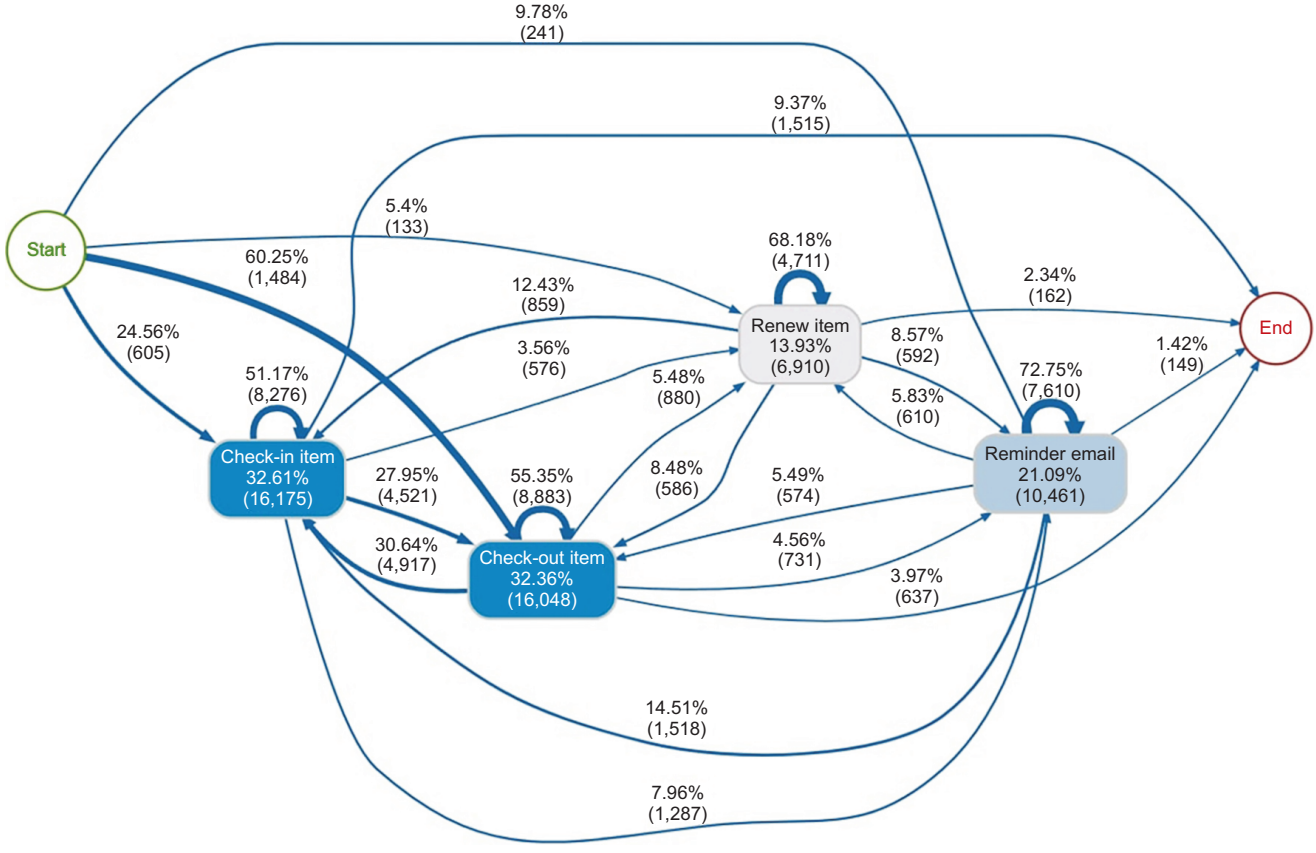


Fig. 7. Overall activity sequence of user interaction with the online library system in 2022 (own data processing).

on library users' interactions when borrowing items from the library in 2022. Note that the number of interactions was slightly lower than in 2018. This is due to the smaller number of library users in 2022 as compared to 2018. Hence, the percentage of proportion is better for representing user interactions with the online library system. Similar to 2018, most library users began their interaction with the system by checking out the item (start → check-out item), accounting for 60.25% of initial interactions in 2022. Unlike in 2018, the library users' check-in and check-out items were much more in number of items, as presented in Fig. 7. There was a strong connection between check-in and check-out items from the library (check-in item → check-out item; check-out item → check-in item). The proportion of e-mail reminders was two times lower than the proportion of reminder e-mails observed in 2018. Meanwhile, the number of renewed items was also higher.

6. DISCUSSION

The exploration of entry and room reservations showed a higher number of users accessing the library in 2018 compared to 2022. However, the number of room reservations in 2018 was much lower than in 2022. This showed that users utilized several learning spaces offered by the library, such as a common space, quiet reading space, learning space, meeting room, and computer room. This aligns with several research studies indicating that the library does not only represent a repository of information but also offers knowledge sharing (De Groote & Scoulas, 2021; Yu et al., 2023), learning resources (Gambles, 2016), a creativity atmosphere (Mahidol University, 2020), and discussion spaces between users (Gambles, 2016; Matichon Public, 2017). Students often visit the university library while waiting for their next class (Unchitti, 2005).

The high number of entries in 2018 reflected that before COVID-19, library users were comfortable using the common learning spaces and library facilities at their convenience. However, the higher numbers of room reservations in 2022 showed that rather than spending time in a common space like before the COVID-19 outbreak, library users preferred to book a limited closed space by booking specific rooms. This is different from the situation in tourism where revenge users surged out after COVID-19 to spend time traveling, compensating for the time they had to spend at home. The library users behaved differently. Users spent time in the library but preferred to book private group learning spaces. One of the possible reflections that can be observed from this finding is that Thai university library users were aware of and continued to take precautions to minimize the risk of being exposed to the virus. This is aligned with the study from Mahidol University (2020), which found that in the post-COVID-19 era, many populations in Thailand have increased their awareness of the virus and infections. They complied and positively responded to the campaigns to reduce and control the spread of the virus.

Among the campaigns promoted by Thailand authorities during COVID-19 were keeping social distance, reducing risk by not being in a crowded space, working in a small group in a closed space with well-air circulation systems, and many other strategies. The post-COVID-19 situation from this study reflects that library users adapted and felt more comfortable using the limited closed collaboration space rather than common space as in pre-COVID-19 situations. As a result, the number of room reservations in 2022 for group meetings within the library has increased accordingly. This is also consistent with Wongmonthar (2021), who found that after the COVID-19 outbreak, people need to be attentive to their health and focus on protecting themselves in order to survive by changing their behavior. By living a life that deviates from the old ways of life by building and adjusting to a new way of living, such as changing behavior by refraining from participating in activities with numerous people, and isolating from places where there are large gatherings or gatherings of people, private and careful use of space has increased.

The most striking finding from this study was the activity of borrowing and returning the resources of library users. Process mining revealed that a large volume of reminder e-mails was sent out prior to COVID-19. A higher number of reminder e-mails is linked to the high volume of library visitors during 2018. Hence, the proportion of events was calculated in relation to all recorded data in that year. Reminder e-mails accounted for 42% of all recorded data in 2018, while only 21% was observed in 2022. Reminder e-mails aimed to remind library users regarding the due date of their checked-out resources. The higher frequency of e-mails indicated the higher numbers of library users who prolonged their return of borrowed items. This shows that library users have neglected to return their books to the library. This may have been influenced by personal traits of library users, such as procrastination behaviors, losing a book, forgetting, or lack of social responsibility. In addition, the reason for higher e-

 https://doi.org/10.1633/JISTaP.2025.13.1.3

mail notifications for users to return books may have been due to the fewer restrictions on visiting the library physically before COVID-19. Hence, library users may not pay attention to the due date as they can easily re-borrow the borrowed resources at any time via the ALIST OPAC system, as well as physically visiting the library.

The smaller proportion of e-mails in 2022 indicated that there was a lower number of e-mails sent to remind of due dates. This showed that users returned books prior to the due date. Hence, no reminder e-mails were sent out in a large number of activities. Many users spent less time holding the borrowed printed resources. This might indicate that users read books or printed resources faster, which resulted in the activity of returning books faster. As claimed by Alomari et al. (2023), based on study on the impact of COVID-19 confinement on reading behavior, it was found that due to the outbreak and quarantine measures, including self-quarantine, physical distancing, prohibition of group activities, school closures, and lockdowns, people changed their behavior. They were more enthusiastic about reading than before, approximately 15% more. They found that people spent time on reading, approximately more than three hours per person. Similarly, Madani et al. (2020) also found that during COVID-19, approximately 15% of participants read more than three hours daily. Apinunmahakul et al. (2023) posited that this may have been due to a change in users' behaviors, because they were accustomed to spending more time by themselves, which resulted in the habit of spending more time reading. In addition, various activities that used to take place and fill time, such as field trips, social gatherings, and traveling to various places before and during the COVID crisis, such as homes, educational institutions, entertainment venues, and tourist attractions, have disappeared. Hence, many people replaced their spare time with reading or other activities that can be performed at home. Hence, the reading capacity has increased.

7. PRACTICAL IMPLICATIONS

Implications for practices derived from this research will help service operations of university libraries, school libraries, public libraries, specialized libraries, and information service institutions to be able to apply them to increase service quality in accordance with the needs of library users. Implications for practices are threefold. Firstly, physical space is still an important element in the library. Regardless of the prior or post COVID-19 effects, university library users still use the library spaces immensely. The physical space needs to be well-managed and well-allocated to suit the needs of library users. The current study suggests that in being aware and taking precautions in 2022, many university library users preferred limited private group working spaces to discuss, share, and collaborate with their peers. The supported systems that allow flexible access and reserving of the learning spaces offered by the university's library have increasingly been used in recent years. Nonetheless, the arrangement and allocation of the library space need to be well-managed. Several previous studies suggested that to support learning activities (Kaewurai & Chaimin, 2019), the library should include:

- Learning space is a space that is designed flexibly to allow users to organize activities according to their own needs and appropriate to their age.
- Inspiration space is an area that opens up the experiences of users who can share their experiences, which can be done through storytelling, displaying work, exhibitions, and training.
- Meeting space is an area that is designed to have both an open system and a closed system. It can be a space to create participatory learning, including formal and informal meetings.
- Space for expression or Performative space is a space to support activities for users to express themselves creatively through various forms of activities and presentations.

Similarly, Andrews et al. (2015) stated that providing learning space can be done by dividing library spaces into different zones. They suggested including activity areas, private space or space for personal study, group workspace, and other types of spaces. Additionally, technology and virtual space can also be used to enhance the library experience. Further exploration of this could shed some light and provide some suggestions for design of the virtual space.

Secondly, this study also provides some suggestions on resource allocation and management. As found in this study, many users still prefer paperbound reading materials, as reflected by the number of library resources borrowed and returned. Based on exploration of the disciplines of the university's library users, it was shown that the greatest number of users were from the social sciences and humanities, political sciences, and Islamic studies. Sufficient resources relevant to these fields are important. The library should consider its regular users and their needs when purchasing new resources to ensure the ef-

ficiency of the library and its resources, both paperbound and digital forms.

Lastly, the results from studying the behavior of library users both before and after the COVID-19 outbreak have resulted in knowledge and in-depth findings. This can be used to inform the designing of information services, improve processes, and serve as a guideline to increase the value of the university library's services or information service institutions in the era. After the COVID-19 outbreak, reading behavior has changed. As supported by several research studies, the users reduced their reading time. Users read more quickly and are more likely to borrow information resources and use the closed areas of the library, such as small meeting rooms and self-learning rooms. This phenomenon can be an effect of users' intentions to return to familiarity and absorb the old atmosphere with things that could not be done during the COVID-19 outbreak. Therefore, libraries need to prepare information resources and services that are ready for use and sufficient to meet needs. In addition, librarians and library staff need to adjust their activity and relevant activities of providing services to be more concise, timely, and faster, such as increasing standards and procedures for inspecting returned resources on the shelves as quickly as possible. Self-boarding and returning service are also possible solutions to increase the process and convenience.

The use of ML and data mining algorithms can also be used in other types of data to discover the insights contained within the trace data. Such findings could be used to provide some insightful recommendations to formulate policies, plans, and strategies for information services. Moreover, the research results can be used for information service institutions that have been affected by the COVID-19 outbreak to readjust, redesign, and reorganize their services to suit the needs of users. This can also help share good practices in data management to improve the library management system further.

8. LIMITATION OF THE STUDY

The dataset used in this study considered data from two years, 2018 and 2022. Hence, some actions might be a sequence of the previous year's activities. For example, library users might borrow items in December of the previous year and return the item in January. However, this study takes all activities equally as long as they happen in 2018 and 2022.

Secondly, this research employed only the automated recorded data. Even though recorded data reflected users' interactions and activities within the system, it does not provide the rationale for user's behaviors. Asking the users about their perceptions through a survey or questionnaire would provide additional information. The questionnaire might also offer some insights into the reason behind users' actions.

Thirdly, the automated record system is subject to electricity and Internet availability. Therefore, some data might be lost when electricity or the Internet malfunction.

Lastly, some external factors might influence the use of library services, such as the persuasiveness of instructors, the number of students in each faculty, the number of librarians, and other variables. However, this study only focused on online behavior regardless of these factors. Additionally, physical behaviors and direct interaction with librarians were outside of this study's scope. Future work should also focus on these factors and examine their influence on library users' behavior.

9. CONCLUSION

This paper explores the behaviors of library users pre and post-global pandemic. The results from the trace data revealed that the number of entries to the university library after COVID-19 was lower than in 2018. However, the number of room reservations increased. This indicated that after COVID-19, users prefer using private and closed spaces. The results from process mining revealed that the time users held borrowed books was shorter than the time they previously held items before COVID-19. One of the posible explanation might be users developing reading skills and spending less time reading. Hence, the sequence of activities from borrowing to returning or renewing the borrowed resources was shorter. This has resulted in reducing the frequency of reminder e-mails sent out by the system. The changes in number of reservations and behaviors suggests that the library needs to take into account the importance of physically closed private spaces. The faster time of returning books indicates the likelihood of borrowing the new resources; hence, librarians need to ensure that the process and activities are able to cope with the speed of the users. The use of ML and data mining algorithms is beneficial. It enables us to unpack the valuable insights hidden within the trace data.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

 https://doi.org/10.1633/JISTaP.2025.13.1.3

REFERENCES

Abdullah, M. N. A. (2021). Revenge tourism: Trend or impact post-pandemic COVID-19? In A. H. G. Kusumah, C. U. Abdullah, D. Turgarini, M. Ruhimat, O. Ridwanudin, & Y. Yuniawati (Eds.), *Promoting creative tourism: Current issues in tourism research* (pp. 623-627). Routledge.

Al-maaitah, T. A., Majali, T., Alsoud, M., & Al-maaitah, D. A. (2021). The impact of COVID-19 on the electronic commerce users behavior. *Journal of Contemporary Issues in Business and Government,* 27(1), 784-793. https://cibgp.com/au/index.php/1323-6903/article/view/599

Alomari, M. A., Khabour, O. F., Alzoubi, K. H., & Aburub, A. (2023). The impact of COVID-19 confinement on reading behavior. *Clinical Practice and Epidemiology in Mental Health,* 19, e174501792304260. https://doi.org/10.2174/17450179-v19-e230505-2022-42

Andrews, C., Wright, S. E., & Raskin, H. (2015). Library learning spaces: Investigating libraries and investing in student feedback. *Journal of Library Administration,* 56(6), 647-672. https://doi.org/10.1080/01930826.2015.1105556

Apinunmahakul, A., Mongsawad, P., & Sudsawasd, S. (2023). Impact of COVID-19 on primary school children's learning. *Journal of Applied Economics and Management Strategy,* 10(1), 51-69. https://kuojs.lib.ku.ac.th/index.php/jems/article/view/4893

Ashikuzzaman, M. (2023). *Traditional library.* https://www.lisedunetwork.com/traditional-library/

Asif, M., & Singh, K. K. (2020). Trends, opportunities and scope of libraries during COVID-19 pandemic. *IP Indian Journal of Library Science and Information Technology,* 5(1), 24-27. https://doi.org/10.18231/j.ijlsit.2020.005

Baharuddin, N. S., & Mohamad Rosman, M. R. (2020). Factors affecting the usage of library e-services in the aftermath of COVID-19 pandemic. *Academic Journal of Business and Social Sciences,* 4(1), 1-14. https://ir.uitm.edu.my/id/eprint/42534/

Ćirić, J., & Ćirić, A. (2021). The impact of the COVID-19 pandemic on digital library usage: A public library case study. *Journal of Web Librarianship,* 15(2), 53-68. https://doi.org/10.1080/19322909.2021.1913465

De Groote, S., & Scoulas, J. M. (2021). Impact of COVID-19 on the use of the academic library. *Reference Services Review,* 49(3/4), 281-301. https://doi.org/10.1108/RSR-07-2021-0043

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review,* 25, 23-38. https://doi.org/10.1016/j.edurev.2018.09.003

Duran, E. (2013). Efficiency in reading comprehension: A comparison of students' competency in reading printed and digital texts. *Educational Research and Reviews,* 8(11), 728-739. https://academicjournals.org/journal/ERR/article-abstract/74855B46066

Dzihni, A. S., Andreswari, R., & Hasibuan, M. A. (2019). Business process analysis and academic information system audit of helpdesk application using genetic algorithms a process mining approach. *Procedia Computer Science,* 161, 903-909. https://doi.org/10.1016/j.procs.2019.11.198

Ehrenpreis, M., & DeLooper, J. P. (2022). *Implementing a chatbot on a library website.* https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1400&context=le_pubs

Gambles, B. (2016). *Future city, future library: Experiences and lessons learned from the library of Birmingham.* https://news.npru.ac.th/u_news/detail.php?news_id=22397&ref_id=LIBRARY

Gao, J., Yang, X., Tang, L., Ren, C., Zhang, X., & Li, Q. (2020). Understanding urban hospital bypass behaviour based on big trace data. *Cities,* 103, 102739. https://doi.org/10.1016/j.cities.2020.102739

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends,* 59(1), 64-71. https://doi.org/10.1007/s11528-014-0822-x

Horstmann, W., & Brase, J. (2016). Libraries and data – Paradigm shifts and challenges. *Bibliothek Forschung und Praxis,* 40(2), 273-277. https://doi.org/10.1515/bfp-2016-0034

Jabr, F. (2013). Why the brain prefers paper. *Scientific American,* 309(5), 48-53. https://doi.org/10.1038/scientificamerican1113-48

Janssenswillen, G., Depaire, B., Swennen, M., Jans, M., & Vanhoof, K. (2019). bupaR: Enabling reproducible business process analysis. *Knowledge-Based Systems,* 163, 927-930. https://doi.org/10.1016/j.knosys.2018.10.018

Joshi, A., & Sadhale, M. (2022). Revenge tourism: An overview of the phenomenon in India. *Multi-Disciplinary Journal,* 1(2), 1-5. http://www.mahratta.org/CurrIssue/2022_March_Spl-1/1.%20Revenge%20Tourism%20an%20overview%20of%20the%20phenomenon%20in%20India-%20Miss.%20Aditi%20Joshi,%20Mrs.%20Manasi%20Sadhale%20.pdf

Kaewurai, R., & Chaimin, C. (2019). [Learning space for digital natives in academic library]. *Journal of Education and Innovation,* 21(4), 366-378. Thai. https://so06.tci-thaijo.org/index.php/edujournal_nu/article/view/152965

Kamolpechara, K., & Soopunyo, W. (2020). Current scenario of university libraries to enhancing lifelong learning. *TLA Research Journal,* 13(1), 80-95. https://so06.tci-thaijo.org/

index.php/tla_research/article/view/243184

Kouzari, E., & Stamelos, I. (2018). Process mining applied on library information systems: A case study. *Library & Information Science Research,* 40(3-4), 245-254. https://doi.org/10.1016/j.lisr.2018.09.006

Kröckel, J., & Bodendorf, F. (2012, July 24-27). Customer tracking and tracing data as a basis for service innovations at the point of sale. In P. Kellenberger (Ed.), *Proceedings of the 2012 Annual SRII Global Conference* (pp. 691-696). IEEE.

MacKenzie, D. (2020). COVID-19 goes global. *New Scientist (1971),* 245(3271), 7. https://doi.org/10.1016/S0262-4079(20)30424-3

Madani, A., Boutebal, S. E., & Bryant, C. R. (2020). The psychological impact of confinement linked to the coronavirus epidemic COVID-19 in Algeria. *International Journal of Environmental Research and Public Health,* 17(10), 3604. https://doi.org/10.3390/ijerph17103604

Mahapornprajak, T. (2021). *COVID accelerates business transformation into digital.* https://www.bot.or.th/th/research-and-publications/articles-and-publications/articles/Article_24Apr2021.html

Mahidol University. (2020). *Behavior of Thai people during COVID-19 become to New Normal.* https://op.mahidol.ac.th/rm/wp-content/uploads/2020/05/Infographic-New-Normal-after-COVID19_May2020.pdf

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research,* 58, 61-68. https://doi.org/10.1016/j.ijer.2012.12.002

Marome, W., & Shaw, R. (2021). COVID-19 response in Thailand and its implications on future preparedness. *International Journal of Environmental Research and Public Health,* 18(3), 1089. https://doi.org/10.3390/ijerph18031089

Matichon Public. (2017). *Thammasat University open creative space for startup district.* https://www.prachachat.net/sd-plus/sdplus-sustainability/news-78749

Millar, M., & Schrier, T. (2015). Digital or printed textbooks: Which do students prefer and why? *Journal of Teaching in Travel & Tourism,* 15(2), 166-185. https://doi.org/10.1080/15313220.2015.1026474

Oyelude, A. A. (2017). Virtual and augmented reality in libraries and the education sector. *Library Hi Tech News,* 34(4), 1-4. https://doi.org/10.1108/LHTN-04-2017-0019

Parlindungan, F., Rahmatillah, R., & Lisyati, L. (2021). Academic reading preferences and behaviors of Indonesian undergraduate students during COVID-19 pandemic. *Tesol International Journal,* 16(4.1), 6-27. https://www.elejournals.com/tij-2021/tij-volume-16-issue-4-1-2021/

Prince of Songkla University Pattani Campus. (n.d.). *Library services for student*. https://www.oas.psu.ac.th/student/

Salvesen, L., & Berg, C. (2021). "Who says I am coping": The emotional affect of New Jersey academic librarians during the COVID-19 pandemic. *Journal of Academic Librarianship,* 47(5), 102422. https://doi.org/10.1016/j.acalib.2021.102422

Sangchantr, R. (2022). Preparedness of library and information services in New Normal: Thai libraries experience. *TLA Bulletin,* 66(1), 21-36. https://so06.tci-thaijo.org/index.php/tla_bulletin/article/view/256267

Srivastava, N., Fan, Y., Rakovic, M., Singh, S., Jovanovic, J., van der Graaf, J., Lim, L., Surendrannair, S., Kilgour, J., Molenaar, I., Bannert, M., Moore, J., & Gasevic, D. (2022, March 21-25). Effects of internal and external conditions on strategies of self-regulated learning: A learning analytics study. *Proceedings of the 12th International Learning Analytics and Knowledge Conference (*pp. 392-403). ACM.

Tosun, N. (2014). A study on reading printed books or e-books: Reasons for student-teachers preferences. *Turkish Online Journal of Educational Technology,* 13(1), 21-28. https://eric.ed.gov/?id=EJ1018172

Unchitti, D. (2005). *Information knowledge in information and searching* (master's thesis). Burapha University, Chonburi, Thailand.

Van Der Aalst, W. (2012). Process mining. *Communications of the ACM,* 55(8), 76-83. https://doi.org/10.1145/2240236.2240257

Wassler, P., & Fan, D. X. F. (2021). A tale of four futures: Tourism academia and COVID-19. *Tourism Management Perspectives,* 38, 100818. https://doi.org/10.1016/j.tmp.2021.100818

Wongmonthar, O. (2021). New Normal: A new way of living-with-COVID 19 culture. *Rusamilae Journal,* 42(1), 47-62. https://so05.tci-thaijo.org/index.php/rusamelae/article/view/251056

Yu, P. Y., Lam, E. T. H., & Chiu, D. K. W. (2023). Operation management of academic libraries in Hong Kong under COVID-19. *Library Hi Tech,* 41(1), 108-129. https://doi.org/10.1108/LHT-10-2021-0342

Zhou, M., & Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. *Learning and Instruction,* 22(6), 413-419. https://doi.org/10.1016/j.learninstruc.2012.03.004

JISTaP https://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 13(1): 51-69, 2025

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2025.13.1.4

Hate Speech Identification and Categorization on Social Media Using Bi-LSTM: An Information Science Perspective

Krishna Kumar Mohbey
Department of Computer Science, Central University of Rajasthan, Ajmer, India
E-mail: kmohbey@curaj.ac.in

Basant Agarwal
Department of Computer Science and Engineering, Central University of Rajasthan, Ajmer, India
E-mail: basant@curaj.ac.in

Nishtha Kesswani*
Department of Data Science & Analytics, Central University of Rajasthan, Ajmer, India
E-mail: nishtha@curaj.ac.in

Maxim Sterjanov
Department of Computer Science, Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus
E-mail: sterjanov@bsuir.by

Yunevich Nikol
Center of Advanced Studies in Digital Development of JSC "Giprosvyaz," Minsk, Belarus
E-mail: yunevich@giprosvjaz.by

Vishnyakova Margarita
Center of Advanced Studies in Digital Development of JSC "Giprosvyaz," Minsk, Belarus
E-mail: vishnyakova@giprosvjaz.by

ABSTRACT

Online social networks empower individuals with limited influence to exert significant control over specific individuals' lives and exploit the anonymity or social disconnect offered by the Internet to engage in harassment. Women are commonly attacked due to the prevalent existence of sexism in our culture. Efforts to detect misogyny have improved, but its subtle and profound nature makes it challenging to diagnose, indicating that statistical methods may not be enough. This research article explores the use of deep learning techniques for the automatic detection of hate speech against women on Twitter. It offers further insights into the practical issues of automating hate speech detection in social media platforms by utilizing the model's capacity to grasp linguistic nuances and context. The results highlight the model's applicability to information science by addressing the expanding need for better retrieval of hazardous content, scalable content moderation, and metadata organization. This work emphasizes content control in the digital ecosystem. The deep learning-based methods discussed improve the retrieval of data connected to hate speech in the context of a digital archive or social media monitoring system, facilitating study in fields including online harassment, policy formation, and social justice campaigning. The findings not only advance the field of natural language processing but also have practical implications for social media platforms, policymakers, and advocacy groups seeking to combat online harassment and foster inclusive digital spaces for women.

Keywords: hate speech detection, social media, deep learning, machine learning, metadata organization, content labelling

Received: September 4, 2024 **Revised:** October 5, 2024
Accepted: November 2, 2024 **Published:** March 30, 2025

***Corresponding Author:** Nishtha Kesswani
https://orcid.org/0000-0002-9044-4160
E-mail: nishtha@curaj.ac.in

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. Open Access articles are automatically archived in the Korea Institute of Science and Technology Information (KISTI)'s Open Access repository (AccessON). All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/). Under this license, the authors retain full ownership of their work, while permitting anyone to use, distribute, and reproduce the content in any medium, as long as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© 2025 [Krishna Kumar Mohbey, Basant Agarwal, Nishtha Kesswani, Maxim Sterjanov, Yunevich Nikol, Vishnyakova Margarita] This article is distributed under the Creative Commons Attribution License (CC BY), allowing unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. It is published by the Korea Institute of Science and Technology Information (KISTI).

1. INTRODUCTION

The rise in popularity of social media platforms has led to a significant increase in the volume of textual information, rendering manual moderation of this content unfeasible (Cao et al., 2020). Social media platforms such as Twitter, Facebook, and Instagram enable users to freely express themselves, which has boosted the proliferation of hate speech and harsh language, thus posing new challenges for information science. Hate speech, especially directed at women, is a major obstacle to creating a secure and welcoming online space. Conventional information retrieval systems usually focus on offering people pertinent content through keyword or semantic search engines. However, disseminating abusive content on digital platforms can pose a significant risk to people, society, governments, and social media platforms (Miškolci et al., 2020). Owing to this, researchers are using powerful computational techniques, such as profound learning, to create automated systems that can identify and reduce hate speech in response to the increasing worry about this issue (Davidson et al., 2017). This study introduces a deep learning (DL) method for automatically detecting hate speech directed against women, specifically on Twitter, aiming to enhance the establishment of a safer and more encouraging online environment. By integrating models that can automatically detect and flag hate speech, this research expands the functionality of retrieval systems. It enhances user safety in digital contexts while also producing higher-quality search results.

1.1. Concept of Hate Speech

Hate speech is significant because it is subjective. According to Fortuna and Nunes (2019), hate speech is rhetoric that targets or belittles groups, instigating violence or hatred based on physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, etc. It may surface in a variety of linguistic approaches, including sharp-witted humor and scathing retorts. Hate speech is potentially detrimental to people and society (Fortuna & Nunes, 2019). Hate speech is prevalent on social media platforms and requires creative methods to tackle this growing problem. Hate speech, which includes abusive words, discriminatory statements, and threats, harms individuals and goes against the values of accessible and respectful communication (Waseem & Hovy, 2016). Targeted hate speech against women is especially worrisome since it encourages gender-based violence and reinforces negative preconceptions and biases.

Detecting hate speech may be challenging due to users' common use of acronyms, slang, and hashtags. Accurate classification of data improves searchability and fosters moral digital curation. For instance, hate speech content needs to be correctly indexed and managed in a public archive that houses millions of tweets to prevent unintentionally spreading damaging viewpoints. The methodology to identify hate speech should enable archivists to identify, separate, and if required, impose access restrictions on particular types of offensive communication. This helps to ensure that sensitive content is used appropriately for research or educational purposes, supporting the ethical management of digital resources. Moreover, the classification methods can be applied to hate speech and other online misconduct, which will help to create complete digital content management systems. Previous research has investigated many methods to detect and address hate speech, including rule-based systems and machine-learning techniques. The changing and developing online language necessitates advanced technologies, leading to the investigation of DL methods.

1.2. Need for the Study

Several research studies have reported the automatic identification of hate speech in benchmark datasets by integrating natural language processing (NLP) with classic machine learning (ML) techniques (Salminen et al., 2020; Watanabe et al., 2018) and DL strategies (Zhang & Luo, 2019). These approaches involve the utilization of metadata, user-based features, and text mining-based features. These features include lexical approaches, grammatical approaches, bag-of-words (BOW), text embedding, sentiment analysis, and others. There is a need to organize metadata surrounding hate speech by creating an effective framework to categorize it and distinguish it from other offensive or neutral language. For academics and politicians alike, it would be essential to organize resources with relevant metadata tags such as "hate speech," "misogyny," or "gender-based harassment" in an online archive or digital library devoted to social media studies. Different models can be incorporated into these systems to automatically classify content, guaranteeing that users can access important information while avoiding undesirable or hazardous content. Such research is expected to improve current hate speech identification methods and guide the creation of better content control tools on social media sites.

Srba et al. (2021) studied hate speech detection from a computer science perspective. They analyzed the contribution of different forms of data, including metadata,

 https://doi.org/10.1633/JISTaP.2025.13.1.4

textual data, videos, and images to hate speech detection. Both models have researched both techniques. However, for DL models to function well, they need a substantial amount of data that has been labeled. Agarwal and Chowdary (2021) have investigated the use of ensemble learning in the context of hate speech identification. Ensemble learning has also demonstrated robust findings. The possible biases of the datasets and algorithms were not considered in these works, even though various contributions have been devoted to analyzing these topics and have provided good classification scores. In the light of the above discussion, the following research hypotheses can be formulated:

H1. The accuracy of hate speech detection on various social media platforms can be greatly enhanced by ML models trained on a range of annotated datasets, decreasing false positives and false negatives.

H2. Utilizing advanced word embeddings in hate speech detection models will enhance the accuracy of classifying nuanced or context-dependent hate speech compared to traditional text representation techniques.

H3. Automated hate speech detection systems improve the organization and retrieval of digital data by effectively classifying harmful content, thus upholding the integrity of information in extensive information systems and databases.

1.3. Objectives

This study is driven by the necessity for an automated and scalable approach to address hate speech directed at women on Twitter. Utilizing DL shows excellent potential by enabling models to acquire intricate patterns and representations from data. DL has been highly successful in NLP tasks, making it well-suited for tackling the issues presented by the ever-changing and context-specific nature of hate speech on social media. The primary objectives of this research are as follows:

- This research offers useful insights into how automated systems might improve the detection and categorization of hate speech on social media sites like Twitter, especially directed toward women, by employing DL models.
- By creating models that can automatically produce precise and context-aware metadata tags, we can enhance the categorization and retrieval of content relevant to hate speech and contribute to effectively curating large-scale social media datasets.
- Rather than concentrating solely on technical measures, we aim to assess the model's performance in detecting hate speech in real-world scenarios, emphasizing the usefulness of accuracy, precision, recall, and F1-score in improving the functioning of content moderation tools.
- To contribute insights into the linguistic features and contextual elements that distinguish hate speech against women on Twitter, thereby advancing our understanding of online gender-based harassment.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature, highlighting the current research on hate speech detection and DL applications in social media. Section 3 outlines the methodology, detailing the dataset collection and preprocessing steps and the architecture and training of the DL model. Section 4 presents the experimental results and discusses the implications of the findings. Finally, Section 5 concludes the paper, summarizing key contributions and discussing potential areas for future investigation.

2. RELATED WORK

The consequences for online safety, social cohesiveness, and freedom of expression have made hate speech detection on social media platforms an essential field of research. Hate speech detection spans multiple aspects. Pérez et al. (2023) have explored a novel method of data curation, sampling, and annotation. In their study, they investigated the impact of contextual information in enhancing hate speech detection. DL's recent advances have greatly improved detection accuracy and scalability, unlike earlier methods that depended on manually created features and conventional ML techniques (Waseem et al., 2017). Hate speech on Twitter frequently takes on gendered forms, according to research; specifically, women are more likely to be the targets of insults, threats, and harassment. To effectively combat sexism and misogyny on social media, it is essential to understand these gender-specific trends (Mocanu et al., 2015). A study by Tontodimamma et al. (2021) investigated the expanse of hate speech using information mapping on the Scopus database. The bibliometric metric was based on co-word methods analysis and relied on different ML algorithms. DL approaches have entirely changed the hate speech detection game. With advancements such as transformer-based models, recurrent neural networks (RNNs), and convolutional neural networks (CNNs), models can learn

intricate representations and patterns from text input. These methods work particularly well for detecting hate speech against women on Twitter because of their greater effectiveness in collecting contextual subtleties and semantic information (Nobata et al., 2016).

Some have voiced worries about biases in hate speech detection algorithms, even if DL-based systems are effective. Unintentionally reinforcing prejudices and worsening existing inequities are the potential outcomes of models trained on biased datasets. Rahman et al. (2021) proposed an information retrieval-based hate speech detection approach involving pooling and active learning methods. The study emphasized task decomposition and annotator rationale techniques. Hate speech detection systems, especially those that target hate speech based on gender, need to be fair and overcome biases, according to experts (Sap et al., 2019). More recent research has looked into novel ways to make hate speech detection algorithms better and easier to understand. The problems presented by the ever-changing types of hate speech on Twitter can be partially addressed by employing domain adaptation approaches, multimodal learning, and adversarial training. Additionally, there is a rising movement to improve the openness and responsibility of hate speech detection systems using explainable AI approaches (Kheddar et al., 2023). Zhang and Luo (2019) explored the long-tail effect in the Twitter dataset and the consequent impact on classifying hate speech. Their study critiques the common practice of micro-averaging, which can bias evaluations towards non-hate classes, and demonstrates improved classification of hate speech in the long tail of datasets.

Identifying abusive language was the primary emphasis of Wiegand et al. (2018)'s research. The authors constructed an abusive lexicon through the use of a variety of traits and lexical resources. Following that, a built lexicon implemented in an support vector machine (SVM) classification was utilized. This study used datasets accessible to the general public (Waseem & Hovy, 2016). It is important to draw attention to the fact that all of the research stated above was conducted using English.

On the other hand, it is worth noting that there have been a few additional studies undertaken in other languages in recent times, including Italian (Del Vigna et al., 2017), German (Köffer et al., 2018), Russian (Andrusyak et al., 2018), and Indonesian (Alfina et al., 2017). Abozinadah et al. (2015) analyzed several machine-learning algorithms to identify offensive tweets written in Arabic. In addition to using three different categorization techniques, they manually identified and labeled five hundred accounts related to the abusive tweets extracted. The naive Bayes (NB) classifier obtained an F1-score of up to 90%, which was the most efficacious. Haidar et al. (2017) proffered a system for detecting and preventing cyberbullying on social media platforms. The authors manually annotated a large dataset of 35,273 tweets from the Middle East region. Through the use of SVM and NB, the authors were able to get the best results in terms of classification, with SVM reaching an F1-score of up to 0.93. Recently, Alakrot et al. (2018) have described creating an inflammatory dataset of Arabic comments on YouTube. After analyzing 150 videos on YouTube, the writers retrieved 167,549 comments from the platform. Sixteen thousand comments were randomly selected for annotation, and three people were responsible for the annotation process.

Waseem and Hovy (2016) employed the logistic regression (LR) classification method to detect racist and sexist content on social media. After manually annotating a dataset of 16,914 tweets, the authors reported that 3,383 had sexist material, 1,972 contained racist content, and 11,559 contained neither sexist nor racist information. The authors used the Twitter application programming interface to extract tweets with a few terms associated with women to generate the dataset. They were able to attain an F1-score of 0.73. Many researchers (Al-Hassan & Al-Dossari, 2019) use this study as a standard of excellence. Pitsilis et al. (2018) propose that to identify instances of racism and sexism in social media, it would be beneficial to use a neural network solution comprising numerous long-short-term-memory (LSTM) based classifiers. In a large number of tests, the authors were able to get the highest possible F1-score of 0.93.

Another group of researchers, Kshirsagar et al. (2018), focused on identifying racism and sexism, and their methodology is likewise based on neural networks. Nevertheless, the authors of this study also utilized word embedding to collect features and merge them with a classifier based on multi-layer perception. A maximum F1-score of 0.71 was reached. Saha et al. (2018) created a methodology to identify instances of hate speech directed towards women. In order to extract features, the authors utilized several methods, including BOW, term frequency-inverse document frequency, and sentence embeddings, along with various classification algorithms, including LR, Extreme Gradient Boosting (XGBoost), and CatBoost. Using the LR classifier, the best F1-score produced was 0.70. A hybrid model that combines CNN and LSTM was suggested by Zhang and Luo (2019) to identify hate speech. Seven datasets were used to apply the authors' method, of

 https://doi.org/10.1633/JISTaP.2025.13.1.4

which five are available to the general public.

In previous years, much progress has been made in utilizing DL to automatically detect hate comments against women on Twitter. The fight against prejudice, for equality, and against new kinds of hate speech is far from over. Research into the fight against cyberbullying and other forms of gender-based violence can progress toward more effective and moral solutions if it builds on the work that has already been done in this area.

3. PROPOSED MODEL

In this section, we discuss the process of hate-speech detection along with each phase. Fig. 1 depicts the entire process schematically. Data in its original form is gathered from different social media sites. The information is preprocessed, labeled, and divided into training and test sets. Before feature extraction, tokenization and padding are used to standardize the text sequences. Word vectors, particularly GloVe, represent words in a dense format. Next, the information goes through classifiers, such as traditional ML models and DL models. In the end, the system categorizes the text into either "Hate," "Offensive," or "Neither."

3.1. Data Preparation

We evaluated hate speech detection using data from Twitter. After applying data preprocessing, it was split into training and testing sets of 80% and 20%, respectively, with a validation split of 20%. The text is categorized as hate speech, offensive language, or neither. It is crucial to remember that this dataset contains content that may be interpreted as racist, sexist, homophobic, or simply offensive.

We undertook a four-stage process to prepare the textual data for the sentiment analysis. Firstly, we extracted the sentiment labels from the data and reorganized the data frame. Numeric labels were then mapped to more descriptive categories, such as "hate speech" or "offensive language," for better comprehension. Second, to address the issue of variable-length text data and to ensure the model's efficacy, we employed text tokenization. This process segments each text into words or subwords. The maximum number of vocabulary words to be considered was set at 20,000. However, tokenization still leaves the issue of variable length unattended. This inconsistency can pose problems for ML models that expect inputs of uniform size. Thus, we pad the shorter sequences with zeros to make the sequences have the same dimensionality. Next, we augment the text representations using GloVe, a pre-trained word embedding model. It helps capture the inherent structure and content within the otherwise latent words and forms the embedding matrix. Words with similar semantic meanings have similar embedding vectors. Finally, as stated earlier, the textual content has been classified into categorical labels. These labels classify the content into distinct classes without stating the natural order. Thus, we use one-hot encoding to better represent the categorical labels. The final one-hot encoded vector represents the relationship between the input and the text classes.

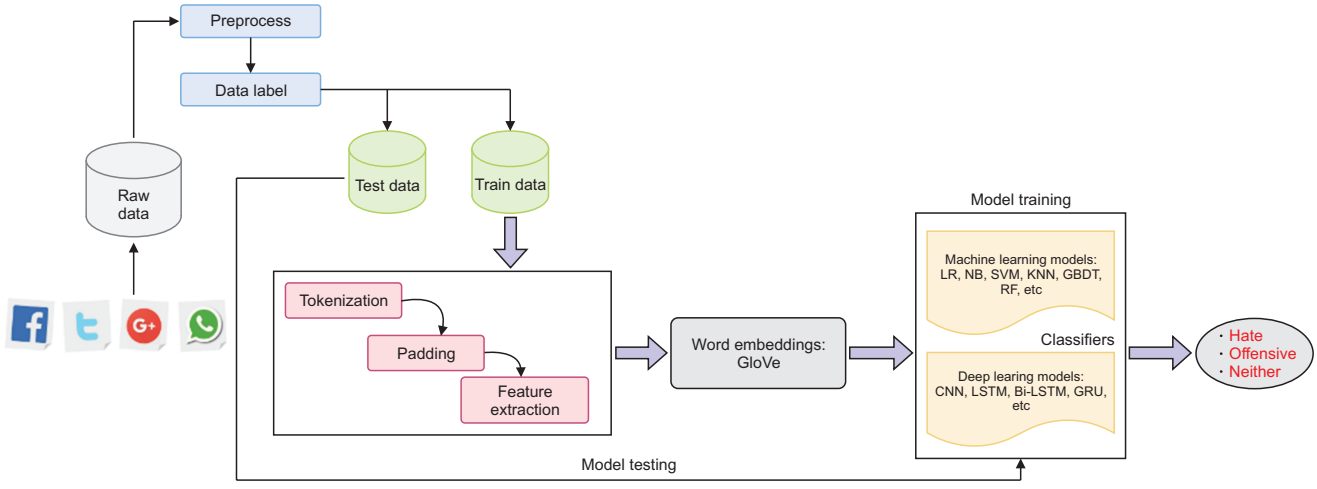


Fig. 1. Schematic framework of hate speech detection using deep learning and machine learning classifiers. LR, logistic regression; NB, naive bayes; SVM, support vector machine; KNN, K-nearest neighbor; GBDT, gradient boosted decision trees; RF, random forest; CNN, convolution neural network; LSTM, long-short-term memory; Bi-LSTM, bidirectional long-short-term memory; GRU, gated recurrent units.

3.2. Classifiers

After applying the data preprocessing, we train different ML and DL classifiers. Here, we discuss each one of them in detail.

3.2.1. Machine Learning Classifiers

Logistic Regression (LR): A fundamental classification algorithm estimates the likelihood that an instance belongs to a specific class. Its high interpretability boosts its extensive use. It is a common choice for binary classification tasks, especially when the relationship between features and the target variable is linear (Sperandei, 2014).

Random Forest (RF): This ensemble learning technique combines several decision trees (DT) to produce correct predictions. It decreases overfitting by averaging predictions from individual trees. It can perform classification and regression tasks and is resistant to outliers and noisy data (Rigatti, 2017).

Support Vector Machine (SVM): SVM is an effective approach for classification and regression. It identifies a hyperplane that best separates various classes while improving their distance. It is suited for high-dimensional spaces and can handle nonlinear interactions with kernel functions (Patle & Chouhan, 2013).

K-Nearest Neighbours (KNN): This classifies occurrences based on the class of their closest neighbors. It is easy to use and requires no training, making it suitable for small datasets. The distance metric selection and the number of neighbors heavily influence its performance (Zhang et al., 2018).

Naive Bayes (NB): A probabilistic algorithm founded on Bayes' theorem. It is especially beneficial for text categorization and other high-dimensional data. Despite its "naive" feature independence assumption, NB is computationally efficient (Berrar, 2019).

Extreme Gradient Boosting (XG Boost): This is a highly efficient version of Gradient Boosting and is well-known for its accelerated performance. Its unique characteristics include a regularization mechanism to check for overfitting and a depth-first strategy for tree pruning to improve generalization capability (Chen & Guestrin, 2016).

3.2.2. Deep Learning Models

Convolution Neural Network (CNN): CNN is a deep neural network that automatically adapts itself to learn the spatial hierarchies of features inherent within the input. CNN consists of convolution layers that use kernels to compute element-wise multiplication on input, aggregate the result to provide a single value, and thus learn the local patterns (Liu et al., 2019).

Bidirectional Long-Short-Term Memory (Bi-LSTM): This is an extension of LSTM that processes input in forward and backward directions using past and future tokens. It has been widely used in sentiment analysis. It uses two separate LSTM layers for processing in different directions. The outputs are concatenated before being used for final predictions (Li et al., 2018).

Long-Short-Term Memory (LSTM): Designed to solve the vanishing gradient issue, LSTM is an extension of RNNs. The input, output, and forget gates are its three gates. They are apt for sequential data processing tasks (Staudemeyer & Morris, 2019).

The reason for selecting a mix of classic ML models and DL models lies in their different strong points, which help tackle different issues in hate speech detection. ML models allow for easier interpretation of results, providing insights into how features contribute to hate speech classification. DL models are suitable for identifying subtle forms of offensive language and hate speech in text by capturing long-term dependencies, sequential patterns, and contextual information that traditional models may struggle to detect. This variety permits a thorough assessment of various approaches, helping the research pinpoint the most effective model or blend of models for tackling issues of hate speech identification.

3.3. Model Architecture

In this section, we discuss detecting Bi-LSTM-based hate speech in detail. Algorithm 1 represents the entire process of hate speech detection.

Algorithm 1: Hate speech detection using Bi-LSTM neural network

Input: Text data (T), Preprocessed labels (L), Maximum sequence length (MAX_LEN), *d*, Word embedding matrix (W)
Output: Sentence class *C={"hate Speech," "Offensive," "Neither"}*

1 Apply preprocessing
  (a) sentence tokenization
  (b) remove punctuations such as " '!()-[]{};:'"\,<>./?@#$%^&*_~" '
  (c) remove stop words
2. for each sentence in *T*
  while sequence length <MAX_LEN do
  (a) Pad the sequence with zeros
  (b) Return padded sequence S_p
  end while
  end for

 https://doi.org/10.1633/JISTaP.2025.13.1.4

3. Lookup_Embeddings(S_p, W)

   Initialize an empty list embedding to store word embeddings

   for each word in the S_p

   (a) Lookup the word's embedding vector from the W matrix

   (b) Append the embedding vector to the embedding list

   (c) Return embeddings list

   end for

4. Initialize hyperparameter for model construction

   learning rate=0.001,
   optimizer='adam',
   epochs=100,
   batch_size=64,
   train_set (T_r)=80%,
   Validation set (T_v)=20%
   test_size (T_s)=20%

5. while each sentence $S \in T_r$ do

   (a) Generate all word embedding vectors in S=[s_1, s_2, s_3, ..., s_n]

   (b) Construct a Bi-LSTM network

   (c) Train the Bi-LSTM model on T_r along with L

   (d) Use a SoftMax classifier to categorize the Bi-LSTM's output as in C

   (e) Validate the model predictions on T_v

   end while

6. for each statement in T_s

   Test the Bi-LSTM model and predict the labels in C

   end for

3.3.1. Forming Word Embeddings

We represent each word in the vocabulary as a dense vector of size d (embedding dimension), which consists of semantic information about the word. Using GLoVE, a pre-trained word embedding matrix w is formed with size *vocabulary size*×d. This matrix associates each word in the vocabulary with a dense vector (embedding). Words with similar meanings tend to have similar embeddings in this vector space. The model learns these pre-trained embeddings during training and may fine-tune them for the specific sentiment analysis task. Each token in the preprocesses sequence is looked up in the embedding matrix. This retrieves its corresponding embedding vector, transforming the text sequence into a sequence of embedding vectors. Fig. 2 provides a visualization of the embedding formed from 100 samples randomly selected. The principal component analysis (PCA) analysis in Fig. 2A depicts that similar words will be close together, but different words will be further apart in meaning or context. In detecting hate speech, this is crucial as words linked to hate speech could gather together, possibly showing offensive language patterns. PCA preserves the overall layout by representing the relationships between distant points and the dispersion of points in the plot, which mirrors the variability in the original multi-dimensional space. The closeness of these words can aid in teaching models to identify offensive language, as they frequently indicate negative feelings or hostility.

Fig. 2B represents a t-distributed stochastic neighbor embedding (t-SNE) plot that offers important insights into how words related to hate speech are distributed relative to neutral words. Words that are closer together on this t-SNE plot probably have similar meanings or are used in the same way. In the context of hate speech, words such as

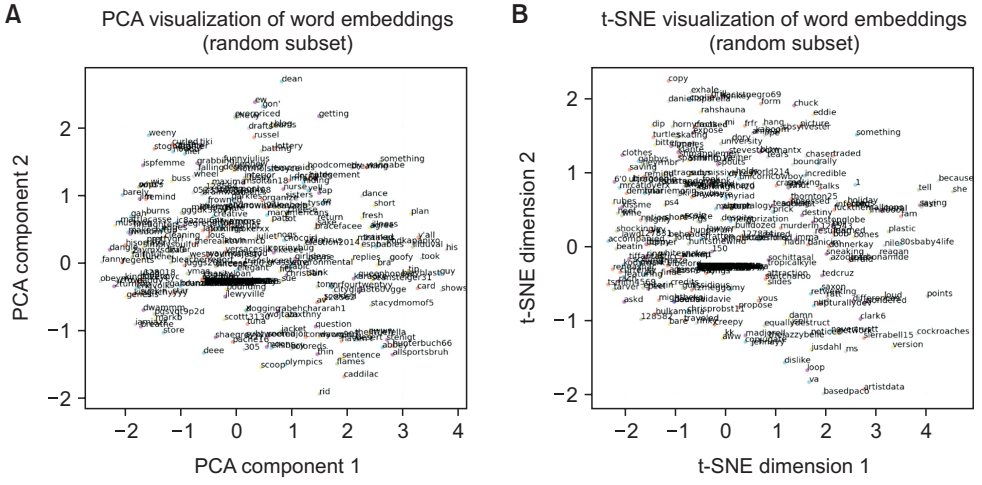


Fig. 2. Visualising embedding: (A) principal component analysis (PCA); (B) t-distributed stochastic neighbor embedding (t-SNE).

racial slurs, offensive terms, and expressions of hate may cluster together. t-SNE emphasizes the preservation of local connections, resulting in compact groupings of related words and facilitating the identification of words with similar meanings or usage contexts.

3.3.2. Bi-LSTM for Hate Speech Detection

In this paper, we propose a Bi-LSTM-based hate speech detection model responsible for capturing the sentiment within the text. Table 1 represents the model summary. It consists of two LSTM layers stacked specially. In forward LSTM, the layer processes the sequence of embedding vectors in a forward direction, capturing the sequential relationships between words from the beginning to the end of the sentence (Equation (1)-(3)). At each time step t in the sequence, the forward LSTM takes the previous hidden state h_{t-1}, the current input embedding x_t from the embedding layer, and a cell state c_{t-1} as input. The forget gate f_t, input gate i_t, and output gate o_t are calculated using sigmoid activation functions.

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f \quad (1)$$

$$i_t = \sigma(W_i \times [h_{(t-1)}, x_t] + b_i) \quad (2)$$

$$o_t = \sigma(W_o \times [h_{(t-1)}, x_t] + b_o) \quad (3)$$

Where σ is the sigmoid function, W_f, W_i, and W_o are the weight matrices, and b_f, b_i, and b_o are the bias vectors.

Table 1. Bidirectional long-short-term memory model summary

Layer (type)	Output shape	Param#
embedding (Embedding)	(None, None, 100)	2,000,000
dropout (Dropout)	(None, None, 100)	0
bidirectional (Bidirectional)	(None, None, 64)	34,048
dropout_1 (Dropout)	(None, None, 64)	0
bidirectional_1 (Bidirectional)	(None, 32)	10,368
dropout_2 (Dropout)	(None, 32)	0
dense (Dense)	(None, 16)	528
dropout_3 (Dropout)	(None, 16)	0
dense_1 (Dense)	(None, 16)	272
dropout_4 (Dropout)	(None, 16)	0
dense_2 (Dense)	(None, 3)	51
Total params: 2,045,267 (7.80 MB) Trainable params: 45,267 (176.82 KB) Non-trainable params: 2,000,000 (7.63 MB)		

c_t is the candidate cell state computed using *tanh* activation function as Equation (4) and updated as Equation (5). Hidden states are updated as Equation (6).

$$c_t = \tanh(W_c \times [h_{t-1}, x_t] + b_c) \quad (4)$$

$$c_t = f_t \times c_{t-1} + i_t \times c_t \quad (5)$$

$$h_t = o_t \times \tanh(c_t) \quad (6)$$

In *backward LSTM*, the layer processes the same sequence of embeddings in a backward direction, understanding the relationships from the end toward the beginning. By combining the outputs of both LSTMs, the model can learn long-term dependencies within the text. This is crucial for sentiment analysis, as the context of surrounding words can influence sentiment. The equations have a similar structure with different weight and bias matrices specific to the backward direction W'_f, W'_i, etc.

3.3.3. Dense Layer

The combined output from the Bi-LSTM is fed into a dense layer with a weight matrix W_d, a bias vector b_d, and a softmax activation function. The softmax function converts the output into a probability distribution over all possible sentiment labels (e.g., "hate speech," "offensive language," "neutral," and "positive"). Each element in the output vector represents the probability of the corresponding sentiment label for the input text sample. The output y is calculated as $y=softmax(W_d \times [h_f, h_b] + b_d)$, where h_f and h_b represent the final hidden states from the forward and backward LSTMs, respectively.

3.3.4. Categorical Crossentropy

Categorical cross-entropy (Ho & Wookey, 2020) is a loss function often employed in multi-class classification situations that combines the projected probability distribution from the Bi-LSTM model with the true probability distribution representing the sentiment labels. Loss is defined as

$$loss = -\sum(y_{true} \times \log(1 - y_{pred}))$$

where y_{true} represents the true label and y_{pred} defines the probability of the same label from the Bi-LSTM model. During training, the Bi-LSTM model seeks to reduce the total category cross-entropy loss across all training samples by modifying the model's weights and biases using an optimization technique.

 https://doi.org/10.1633/JISTaP.2025.13.1.4

4. EXPERIMENTATION AND RESULTS

4.1. Dataset Gathering and Collection Process

The dataset used in this study is provided by Davidson et al. (2017). The data compiled by *Hatebase.org* was searched for tweets with *hate speech* lexicon. These tweets were randomly sampled to form a 25k tweet dataset that CrowdFlower workers manually coded. The tweets were in three categories: hate speech, offensive but not hate speech, or neither offensive nor hate speech. The categorization did not solely rely on the words but also considered their context. Three or more people labeled each tweet, and the final decision was made based on a majority decision. Some tweets were not labeled because of absentia of a majority decision. The process resulted in 24,802 tweets, of which 5% were labeled as hate speech, with 76% at 2/3 measure and 53% meeting a 3/3 measure of offensive language. The remainder were non-offensive. Fig. 3 depicts the statistics of the dataset.

4.2. Experimental Setup

With a 64-bit Windows operating system installed, the experiment was conducted on an Intel(R) Core (TM) i7-6700 CPU @ 3.40GHz 3.41 GHz processor. It has a 16 GB memory capacity. In Python 3.7, the Jupyter Notebook environment was used to program the code. Different libraries were used, such as sci-kit-learn for model analysis and computing performance, pandas for data manipulation and visualization, and numpy for computing operations. We have implemented TensorFlow 2.10, a DL framework. The Natural Language Toolkit Python package with punkt tokenizer was utilized to remove stop words.

4.3. Hyperparameter Tuning

For experimentation purposes, we used batch size 64 and categorical cross-entropy for loss, and each model was optimized by Adam optimizer. We have used early stopping with patience 5 on validation accuracy to prevent overfitting. The embedding dimensions were set to 100. We selected a maximum of 20,000 features and 512 as text length to be considered on each tweet.

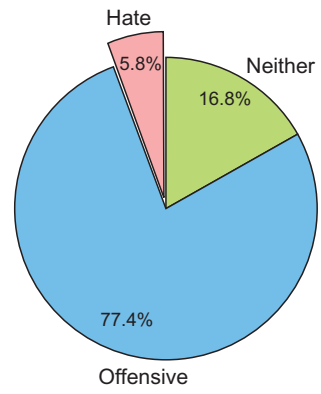
4.4. Evaluation Technique

The model's performance was evaluated based on precision, recall, and F1-score, as discussed in Equation (7)-(9). Besides these attributes, ML models have also been evaluated on micro and macro averages for precision, recall, and F1-score. The micro-average metrics are calculated by adding true-positive $True_{pos}$, $False_{Neg}$, and $False_{pos}$ occurrences of hate speech tweets in the predictions, independent of the instances of the classes. The macro average is the precision, recall, and F1-score average for different classes. As already stated, the dataset is imbalanced, and there are more non-hate speech tweets than hate speech tweets; thus, micro-averaging cannot depict the results of the minority class. We have also constructed a confusion matrix that records the number of *True Positives* ($True_{pos}$), *True Negatives* ($True_{Neg}$), *False Positives* ($False_{pos}$), and *False Negatives* ($False_{Neg}$) and in turn, represents the actual and predicted positives and negatives. Besides this, the qualitative analysis of performance has also been evaluated on accuracy measures as depicted in Equation (10).

a. Precision (P): The ratio of tweets classified as hate speech among the total retrieved tweets as represented by Equation (7).

$$Precision = \frac{True_{pos}}{True_{pos}+False_{pos}} \quad (7)$$

A



B

Category	Label	Split	Count
Hate speech	0	Test	286
	0	Train	1,030
	0	Val	114
Neither	2	Test	833
	2	Train	2,997
	2	Val	333
Offensive language	1	Test	3,838
	1	Train	13,816
	1	Val	1,536

Fig. 3. Dataset statistics: (A) Original dataset; (B) dataset after split.

b. Recall (R): This is the ratio of tweets defined truly as hate speech to the total number of hate speech tweets in ground truth (Equation (8)).

$$Recall = \frac{True_{pos}}{True_{pos}+False_{Neg}} \quad (8)$$

c. F1-Score (F1-score): F1-score is the harmonic mean of P and R depicted by Equation (9).

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (9)$$

d. The confusion matrix comprises the following:

1. *True Positives* ($True_{pos}$): For a data point to be considered positive, its actual class and the prediction must be true.

2. *True Negatives* ($True_{Neg}$): For data points, a true negative occurs when the actual class and the forecast are incorrect.

3. *False Positives* ($False_{pos}$): A false positive is called a false positive when a data point's genuine prediction is based on an incorrect class.

4. *False Negatives* ($False_{Neg}$): An example of a false negative would be a data point whose real class contradicts the prediction.

Table 2 represents the various cells in the confusion matrix. The predicted positives and negatives are the predicted values by the models against actual values. A high $True_{pos}$ indicates the high performance of the model, while a high $False_{Neg}$ suggests that the model missed many instances of hate speech. Similarly, a high $False_{pos}$ means incorrect flagging of non-hate speech content, and high $True_{Neg}$ values indicate that the model is correctly identifying non-hate speech.

e. Accuracy: This is the ratio of the total number of true predictions (positives and negatives) to total predictions (true and false positives and negatives), as represented in Equation (10).

Table 2. Confusion matrix model

	Predicted positive	Predicted negative
Actually positive	*True Positive* ($True_{pos}$)	*False Negative* ($False_{Neg}$)
Actually negative	*False Positive* ($False_{pos}$)	*True Negative* ($True_{Neg}$)

f.

$$Accuracy = \frac{True_{pos}+True_{Neg}}{True_{pos}+True_{Neg}+False_{pos}+False_{Neg}} \quad (10)$$

4.5. Result Analysis

XG Boost emerged as the best performer among the ML variants. It scored an outstanding precision (class 0) of 99.19%, suggesting a high level of accuracy in accurately detecting non-hate speech occurrences, and a recall (class 0) of 97.22%, effectively catching the majority of genuine non-hate speech instances. For hate speech (class 1), the model attained a precision of 18.64%, indicating that it correctly predicts hate speech 18.64% of the time. The recall (class 1) for hate speech was 28.18%, implying that it correctly detected 28.18% of actual hate speech cases. DT demonstrated great accuracy but needs improvement in class 1 recall (30.69%). It struggled to catch every incident of hate speech. LR fared well for non-hate speech (class 0), with 98.63% precision and 96.92% recall. However, its performance in hate speech (class 1) was unsatisfactory, with low precision (17.92%) and recall (25.45%). RF, like LR, excelled at class 0 precision and recall. For hate speech (class 1), RF had a 23.66% precision and a 30.21% recall. SVM had the maximum precision for non-hate speech (class 0) at 99.38%. However, its recall for hate speech (class 1) was extremely low at 18.29%, implying that it missed many actual hate speech incidents. XG Boost exhibited a balanced approach with high accuracy (99.19%) for non-hate speech (class 0). Hate speech had a moderate recall rate (28.18%) in class 1. The F1-score reflects this balance, making XG Boost an attractive option. Table 3 represents the comparative results of different ML models.

The DL models were evaluated across training, validation, and test data sets, and their effectiveness was determined. Table 4 shows how well three models—CNN, LSTM, and Bi-LSTM—performed on various data partitions (train, validation, and test) using four metrics: accuracy, precision, recall, and F1-score. During training, the CNN model attained a notable accuracy of 96.48%. However, its performance declined on the validation set with an accuracy of about 85.35% on test sets, with precision, recall, and F1-score also indicating a similar decrease. The LSTM model showed good performance on the training set, achieving 94.92% accuracy, and continued to exhibit strong performance on the validation (89.18%) and test (89.15%) sets. LSTM achieved high precision, recall, and F1-score, with a slight decrease between training and testing. Consistently, the Bi-LSTM model surpassed both CNN and LSTM models, achieving the highest training

 https://doi.org/10.1633/JISTaP.2025.13.1.4

Table 3. Performance comparison of different ML models

Model	Class	Accuracy	Precision	Recall	F1-score
NB	0.0		0.956289461	0.42090637	0.584533175
	1.0		0.065217391	0.677419355	0.11898017
	Overall	0.435343958			
	Macro average		0.510753426	0.549162863	0.351756672
	Weighted avgerage		0.906136322	0.435343958	0.55832997
DT	0.0		0.959508938	0.952330056	0.955906019
	1.0		0.289808917	0.326164875	0.306913997
	Overall	0.917086948			
	Macro average		0.624658928	0.639247465	0.631410008
	Weighted avgerage		0.921815514	0.917086948	0.919378124
KNN	0.0		0.958018472	0.975630611	0.966744334
	1.0		0.409326425	0.283154122	0.334745763
	Overall	0.936655235			
	Macro average		0.683672448	0.629392367	0.650745048
	Weighted avgerage		0.927135865	0.936655235	0.931172899
LR	0.0		0.952715259	0.98631894	0.969225922
	1.0		0.438596491	0.17921147	0.254452926
	Overall	0.940891668			
	Macro average		0.695655875	0.582765205	0.611839424
	Weighted avgerage		0.923778576	0.940891668	0.928995608
RF	0.0		0.955615753	0.980333476	0.96781682
	1.0		0.417721519	0.23655914	0.302059497
	Overall	0.938470849			
	Macro average		0.686668636	0.608446308	0.634938158
	Weighted avgerage		0.925340891	0.938470849	0.930345306
SVM	0.0		0.949356749	0.99380077	0.971070496
	1.0		0.516666667	0.111111111	0.182890855
	Overall	0.944119427			
	Macro average		0.733011708	0.55245594	0.576980676
	Weighted avgerage		0.925003202	0.944119427	0.926708559
XG Boost	0.0		0.953359359	0.99187687	0.972236773
	1.0		0.577777778	0.186379928	0.281842818
	Overall	0.946540246			
	Macro average		0.765568568	0.589128399	0.627039796
	Weighted avgerage		0.932220109	0.946540246	0.93337861

NB, naive Bayes; DT, decision tree; KNN, K-nearest neighbours; LR, logistic regression; RF, random forest; SVM, support vector machine; XG Boost, Extreme Gradient Boosting.

Table 4. Performance comparison of deep learning models on train, validation, and test set, respectively

Model	Split	Accuracy	Precision	Recall	F1-score
CNN	Train	0.9648	0.9672	0.9639	0.9655
	Validation	0.8535	0.8549	0.8513	0.8513
	Test	0.85031	0.84908	0.85164	0.85034
LSTM	Train	0.9492	0.9522	0.9457	0.9489
	Validation	0.8918	0.8946	0.8903	0.8924
	Test	0.89146	0.888243	0.894047	0.891078
Bi-LSTM	Train	0.9651	0.9674	0.9632	0.9653
	Validation	0.8901	0.8927	0.8886	0.8906
	Test	0.89449	0.892456	0.896078	0.894235

CNN, convolutional neural network; LSTM, long-short-term memory; Bi-LSTM, bidirectional long-short-term memory.

accuracy of 96.51%, impressive validation accuracy of 89.01%, and test accuracy of 89.45%. The Bi-LSTM model had the highest precision, recall, and F1-score values, showing balanced performance and an increase in recall (89.61%) on the test set, indicating its superior ability to detect more true positives than other models.

From the results, we can deduce that CNN achieved competing results. Its high accuracy means that it correctly distinguishes a large proportion of hate speech and non-hate speech incidents. However, we find a trade-off between precision and recall. While the precision for non-hate speech is high, the recall for hate speech is lower. This shows that while CNN is good at identifying non-hate speech, it may overlook some cases of hate speech. LSTM regularly outperforms across all criteria. Its balanced precision and recall for both classes show that it balances erroneous positives and false negatives. However, LSTM's accuracy is slightly lower than that of CNN. Table 4 represents the performance of different DL models.

The performance of different ML-based classifiers can be visualized using the confusion matrix. From Fig. 4, we can deduce that the DT model correctly predicted TP instances (true positives) and TN instances (true negatives). It made some FP errors (false positives) and a few FN errors (false negatives). Overall, the DT model shows a balanced performance. The KNN model achieved high TP and TN counts and had minimal FP and FN errors. Thus, it exhibits robust classification. The LR model performed well regarding TP and TN but had a moderate number of FP and FN errors. The LR model is reliable but not perfect. The NB model had a high TN count but relatively low TP, and it made several FP errors. Thus, it struggles with sensitivity. The RF model excelled in both TP and TN and had minimal FP and FN errors. Thus, it turns out to be a potential second-best performer. The SVM model achieved high TN but relatively low TP and can be tuned for better sensitivity. The XG Boost model had the highest TP count and made few FP and FN errors. Thus, the XG Boost model is the most accurate among the models.

Fig. 5 depicts the confusion matrices for DL models. The LSTM accurately predicted 95 instances of label 0 but misclassified 112 instances of label 1 as label 0, and 16 cases of label 2 as label 0. It correctly identified label 1 173 times while incorrectly predicting label 0 101 times. It exhibited a higher accuracy for label 2, with 698 correct predictions. Bi-LSTM accurately recognized label 0 in 163 instances but misidentified it as label 1 108 times and label 2 2 times. Label 1 had 115 valid identifications, while label 2 had 693 appropriate classifications. CNN outperformed LSTM and Bi-LSTM, with 216 correct predictions. However, its accuracy for label 0 (53 correct identifications) was lower than LSTM (95) and Bi-LSTM (163). For label 2, CNN made 541 right predictions. Thus, LSTM excelled at predicting label 2, whereas Bi-LSTM performed very well for both label 2 and label 1. CNN outperformed the other algorithms in identifying label 1 but struggled with label 0. Differences in performance could be related to architectural variances and training information.

Fig. 6 shows the variation in training and loss across epochs for CNN, LSTM, and Bi-LSTM models. CNN's training loss lowers steadily as the number of epochs grows. This shows that the model is effectively learning from the training data. CNN's training accuracy gradually improves, indicating that it accurately predicts more instances as training advances. The training loss of LSTMs diminishes with epochs; however, it is more pronounced

 https://doi.org/10.1633/JISTaP.2025.13.1.4

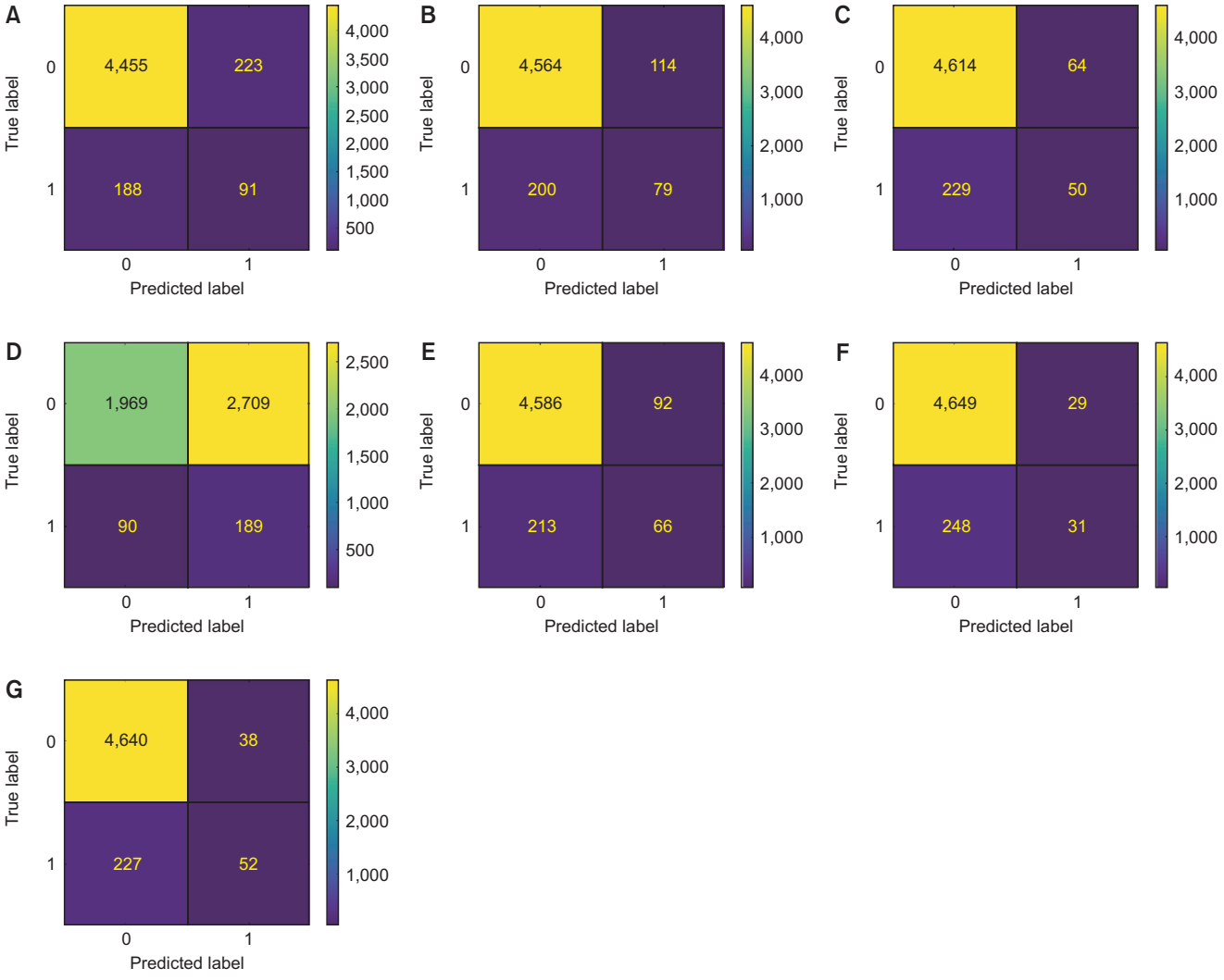


Fig. 4. Confusion matrices for machine learning models: (A) Decision tree; (B) K-nearest neighbours; (C) logistic regression; (D) naive Bayes; (E) random forest; (F) support vector machine; (G) Extreme Gradient Boosting.

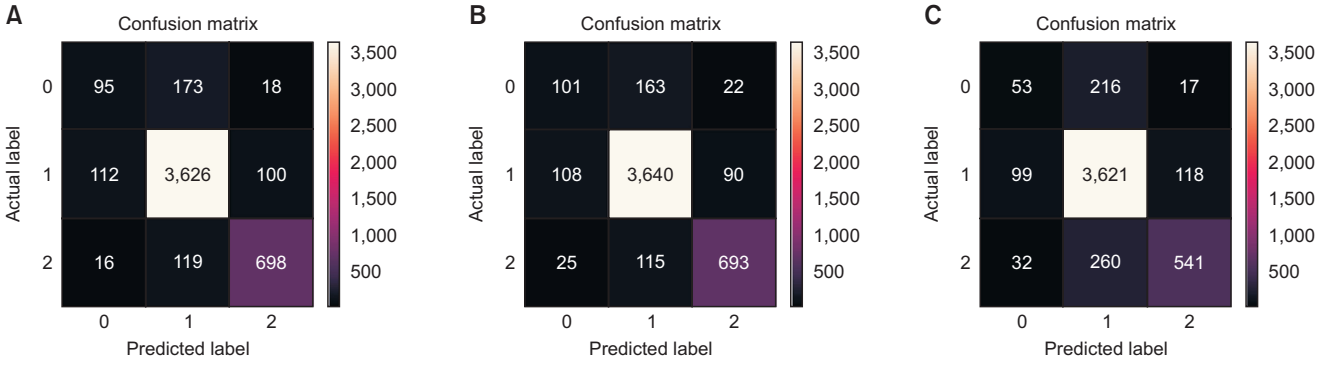


Fig. 5. Confusion matrix of deep learning models: (A) Long-short-term-memory; (B) bidirectional long-short-term memory; (C) convolutional neural network.

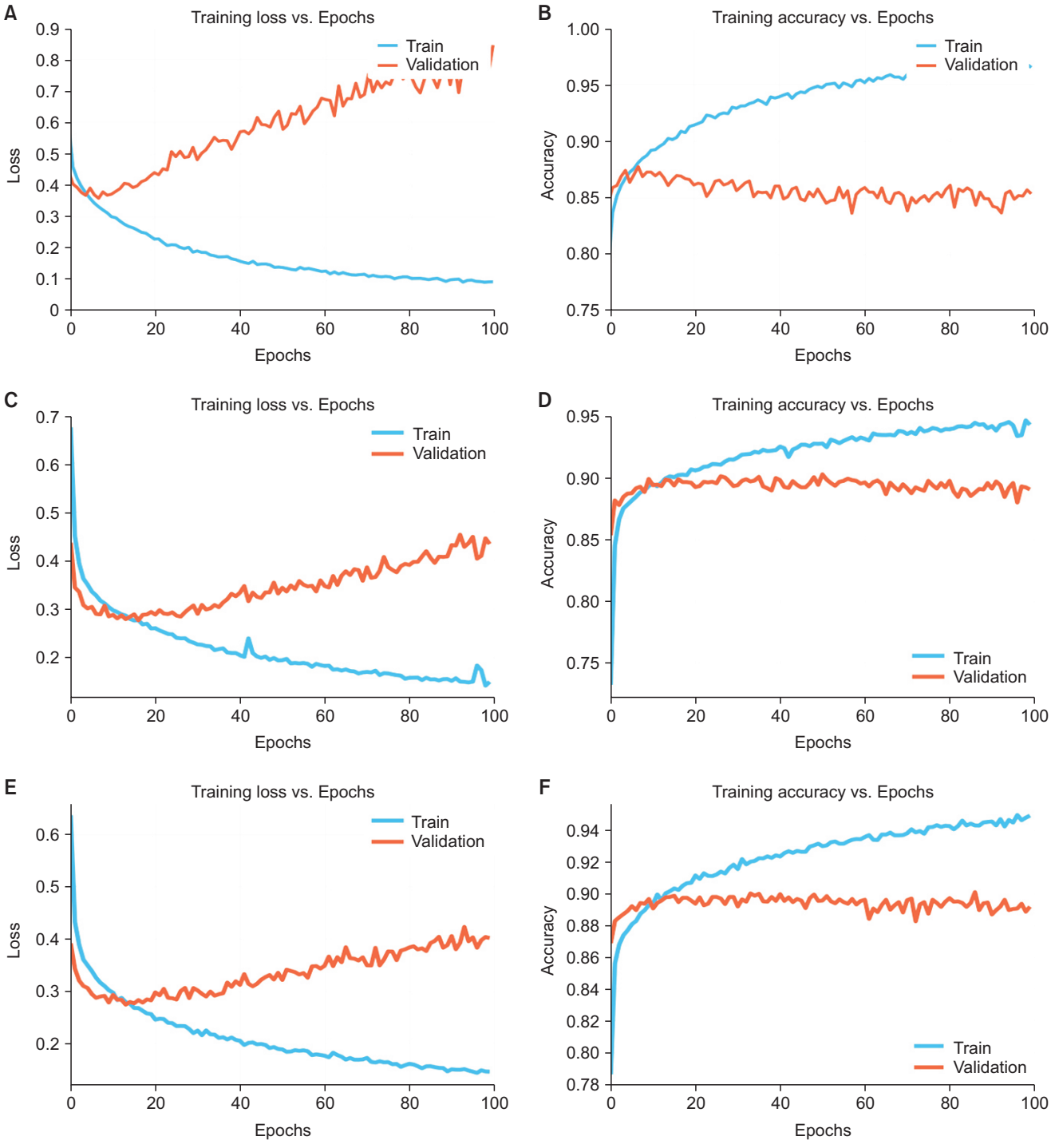


Fig. 6. Training and loss variation with epochs: (A), (B) convolutional neural network; (C), (D) long-short-term-memory; (E), (F) bidirectional long-short-term memory.

than that of CNNs. This indicates that the LSTM is learning efficiently. LSTM training accuracy improves similarly to CNN, albeit with a slightly steeper gradient. Bi-LSTM's training loss is rapidly decreasing, indicating good learning. It converges faster than either CNN or LSTM. Bi-LSTM's training accuracy rapidly increases, outperforming the other models. It reaches high accuracy. During training, the Bi-LSTM model outperforms other models

 https://doi.org/10.1633/JISTaP.2025.13.1.4

in terms of accuracy and loss reduction. LSTM follows closely, with CNN trailing significantly.

4.6. Discussion

This experiment examined the efficacy of various DL architectures for detecting hate speech in text data. We evaluated the performance of a CNN, LSTM, and Bi-LSTM model. To isolate the impact of the architecture, all models were trained using the same hyperparameters. The Bi-LSTM model outperformed the CNN and LSTM algorithms when assessing hate speech. This shows that Bi-LSTM's capacity to identify long-term dependencies in both directions of text sequences may be critical for accurately detecting hate speech.

CNN underperformed compared to Bi-LSTM in hate speech detection. CNNs handle text data with convolutional filters that have a fixed window size. This window captures local word associations but may struggle with hate speech that relies on context dispersed throughout the text. Sarcasm or implied hate speech, for example, may necessitate understanding the overall sentiment of the sentence rather than just a few adjacent words. Standard CNNs analyze text in a single direction (often left to right). Hate speech frequently uses subtle wordings and word placement to communicate its vile meaning. Bi-LSTMs, which can evaluate text in both directions, can better capture these sentence-level order-dependent correlations.

This research enhances metadata tagging systems and information retrieval frameworks by ensuring that the Bi-LSTM model can detect subtle linguistic patterns, including slang and implicit bias. This work shows how ML models can be included in more comprehensive content organization systems, especially in the information science domain, going beyond recall and precision metrics. The model's capacity to differentiate between neutral content, hate speech, and inflammatory language has immediate applications for social media platforms and digital archives looking to improve their metadata systems. These findings highlight the difficulties in detecting hate speech in the real world, including language diversity, irony, and the dynamic nature of online debate. This model provides a scalable approach to these problems by detecting hate speech and classifying it to facilitate content retrieval, analysis, and moral moderation.

Fig. 7 compares the accuracy of hate speech detection on ML and DL models. This distribution highlights the trade-offs between simplicity and performance in ML models. While ML approaches are interpretable and computationally efficient, they struggle with noisy social media data because of their simple feature representations and inability to handle context. DL models use neural networks to learn relevant features from raw text data, making them capable of capturing complicated patterns and contexts. They excel at comprehending context, processing loud text, and detecting subtle linguistic signs. Looking at it from the viewpoint of information science, this automation does not just make content moderation more efficient, but boosts information retrieval systems. Organizing content by how harmful it is, these models help improve search results and suggestions, lowering the likelihood of users coming across dangerous material. Furthermore, the capability to identify and categorize hate speech guarantees that data stays compliant with ethical

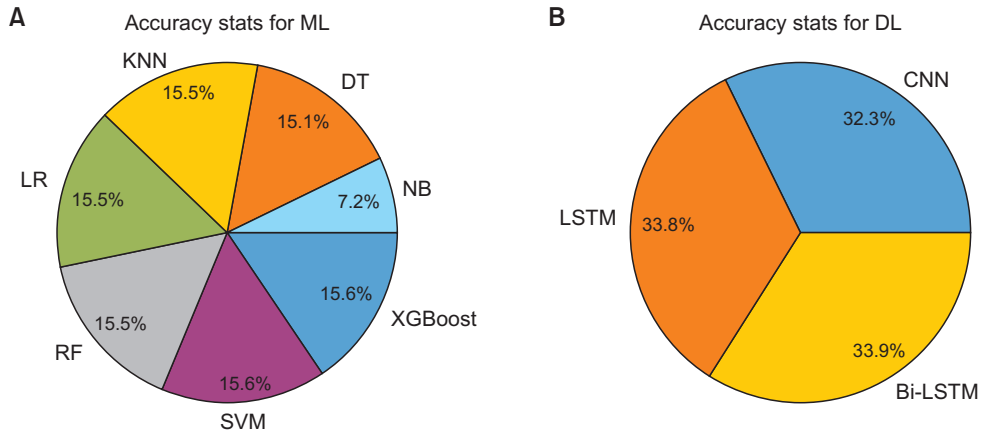


Fig. 7. Accuracy comparison for machine learning (ML) and DL methods. KNN, K-nearest neighbours; DT, decision tree; NB, naive Bayes; XG Boost, Extreme Gradient Boosting; SVM, support vector machine; RF, random forest; LR, logistic regression; CNN, convolutional neural network; Bi-LSTM, bidirectional long-short-term memory; LSTM, long-short-term memory.

and legal norms, contributing to upholding the integrity and reliability of the data in these systems. Therefore, hate speech detection based on ML is in line with the objectives of information science, promoting the creation of secure and structured digital spaces while ensuring data integrity and fairness.

5. FUTURE STUDIES

Several challenges in hate speech detection plague information systems, as listed below:

5.1. Cultural Variability

Each culture's definition of hate speech is influenced by its distinctive linguistic nuances, social norms, and historical contexts. This diversity requires the creation of flexible detection systems that can identify and understand culturally unique phrases and insults. For example, a term considered offensive in one culture could be deemed appropriate in another, resulting in potential misidentifications. Moreover, with the development of online communication, different slang and references constantly appear, often differing greatly depending on the community. Information science needs to focus on context-aware algorithms that consider cultural factors to improve accuracy and relevance in hate speech detection. In the end, focusing on cultural differences will enhance the efficiency of information retrieval systems and help create fairer digital spaces.

5.2. Evolving Nature of Online Language

The changing online language greatly affects the detection of hate speech, especially in information retrieval systems. Detecting offensive language is made difficult by the emergence of new slang, acronyms, and rapidly evolving terminologies on social media platforms such as Twitter and Facebook. Models must constantly adjust to new expressions and cultural contexts to keep up with this dynamic evolution, as once harmless language may now be linked to hate speech. Moreover, the complexity is increased by multilingualism and code-switching, which involve mixing languages, making it challenging for conventional keyword-based systems. To tackle this issue, information retrieval systems must include real-time updates and models aware of context. Utilizing methods like transfer learning and cross-linguistic training aids in maintaining the effectiveness of these systems. Adjusting to the constantly changing nature of online language helps categorize harmful content more accurately and enhance user experiences.

5.3. Transfer Learning Across Languages

In information science, transfer learning is crucial for surpassing language obstacles and improving model generalization. Transfer learning enables efficient identification of hate speech in various linguistic environments by utilizing pre-trained models from well-resourced languages like English and adjusting them with data from limited resources. This is especially beneficial because hate speech can manifest uniquely depending on cultural and regional influences in different languages. It helps adjust detection systems for poorly represented languages, ensuring models are more inclusive and can be used worldwide. Furthermore, exposing hate speech classifiers to different linguistic structures and expressions can enhance their robustness through cross-linguistic transfer. This aligns with information science objectives, aimed at creating more accessible, equitable, and effective systems for retrieving information and moderating content. Future research can concentrate on creating multilingual datasets and use of transfer learning methods to improve the scalability and efficiency of hate speech detection in various languages.

5.4. Lack of Annotated Data

It is often difficult to find high-quality, annotated hate speech datasets in many languages, particularly for languages that are not well-represented. Insufficient training data may limit the model's generalization capacity across different languages. A model trained on English data might not work effectively in languages with fewer resources, where the labeling of offensive material may be scarce or unreliable. Hence, it is a potential future direction to work on expanding the annotated data.

5.5. Ethical Concerns

Furthermore, addressing the ethical implications of automated hate speech identification, such as privacy problems and the possibility of censorship, is critical. Incorporating user feedback loops can increase the system's accuracy over time, transforming it into a dynamic tool for developing methods against hate speech.

6. CONCLUSION

In social media, the proliferation of hate speech has become a pressing issue, necessitating the development of sophisticated algorithms capable of accurately detecting and mitigating such content. The comparative analysis of

 https://doi.org/10.1633/JISTaP.2025.13.1.4

ML models for tweet hate speech detection revealed that DL models, particularly the Bi-LSTM network, offer superior performance over traditional ML approaches. The Bi-LSTM model's proficiency stems from its unique architecture, which processes data sequences in both forward and backward directions, thereby capturing contextual information from all temporal dependencies within the text. This bidirectional processing is particularly advantageous for understanding the nuanced semantics and complex structures of natural language, which are characteristic of social media communication.

Several metrics have been considered when assessing the performance of various models. The accuracy of the Bi-LSTM model in classifying tweets as hate speech or non-hate speech has consistently outperformed other models. This high accuracy rate indicates the model's ability to discern subtle linguistic cues and patterns that distinguish hate speech from benign expressions. By enabling digital libraries, social media platforms, and other information systems to better categorize hate speech, the Bi-LSTM model created in this study contributes to the ethical management and general accessibility of digital content. This research emphasizes the valuable effects of these technologies on content management and retrieval. As the landscape of online discourse continues to change, it provides a flexible and scalable strategy that is in keeping with the larger objectives of information science.

Moreover, the training and validation loss graphs for the Bi-LSTM model demonstrate a rapid convergence to a lower loss value, suggesting that the model is learning effectively and generalizing well to unseen data. This is further corroborated by the stability of the learning curve, which exhibits minimal fluctuations in accuracy and loss across epochs, indicating a robust model less prone to overfitting. The efficiency of the Bi-LSTM model is also noteworthy. Despite the inherent complexity of processing natural language data, the model's architecture enables it to handle large volumes of text while maintaining high computational efficiency. This is crucial for real-time applications where timely detection of hate speech is essential. By leveraging the Bi-LSTM model, social media platforms can filter out hate speech more effectively, creating a safer and more inclusive online environment. This enhances the user experience and aligns with the ethical and legal standards governing digital communication. As the digital landscape continues to evolve, the Bi-LSTM model's adaptability and scalability will be instrumental in addressing the challenges of moderating content on social media platforms.

Future research in hate speech detection with DL could greatly benefit from advances in transformer-based models such as bidirectional encoder representations from transformers, which have demonstrated exceptional success in detecting context and nuances in text. These models might be fine-tuned to detect hate speech, improve accuracy, and lower false positives. Another area of focus could be creating multimodal models that consider not only textual content but also images, videos, and user metadata. This holistic approach may provide a more complete comprehension of the content and its intent. Furthermore, creating larger and more diverse datasets encompassing numerous modes of communication, platforms, and languages would aid in developing more robust models. Collaboration with social media networks to access real-time data could improve the models' usefulness.

CONFLICTS OF INTEREST

The authors of this manuscript state that they have no conflict of interest.

ACKNOWLEDGMENTS

This work was supported by the Department of Science and Technology India, International Cooperation Division (Grant No: DST/INT/BLR/P-38/2023(G)) and the Belarusian Republican Foundation for Fundamental Research (Agreement No. F23INDG-011).

REFERENCES

Abozinadah, E. A., Mbaziira, A. V., & Jones, J. H., Jr. (2015). Detection of abusive accounts with Arabic Tweets. *International Journal of Knowledge Engineering,* 1(2), 113-119. https://www.ijke.org/show-36-54-1.html

Agarwal, S., & Chowdary, C. R. (2021). Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. *Expert Systems with Applications,* 185, 115632. https://doi.org/10.1016/j.eswa.2021.115632

Alakrot, A., Murray, L., & Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in Arabic. *Procedia Computer Science,* 142, 174-181. https://doi.org/10.1016/j.procs.2018.10.473

Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017, October 28-29). Hate speech detection in the Indonesian language: A dataset and preliminary study. *Proceedings of the 2017 International Conference on Advanced Computer Science*

and Information Systems (pp. 233-238). IEEE.

Al-Hassan, A., & Al-Dossari, H. (2019, February 23-24). Detection of hate speech in social networks: A survey on multilingual corpus. In D. Nagamalai, & D. C. Wyld (Eds.), *Proceedings of the 6th International Conference on Computer Science and Information Technology* (pp. 83-100). CS & IT-CSCP.

Andrusyak, B., Rimel, M., & Kern, R. (2018, December 7-9). Detection of abusive speech for mixed sociolects of Russian and Ukrainian languages. In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proceedings of the 12th Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2018)* (pp. 77-84). Tribun EU.

Berrar, D. (2019). Bayes' theorem and naive bayes classifier. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (Vol. 1, pp. 403-412). Elsevier.

Cao, R., Lee, R. K. W., & Hoang, T. A. (2020, July 6-10). DeepHate: Hate speech detection via multi-faceted text representations. *Proceedings of the 12th ACM Conference on Web Science* (pp. 11-20). ACM.

Chen, T., & Guestrin, C. (2016, August 13-17). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media,* 11(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017, January 17-20). Hate me, hate me not: Hate speech detection on Facebook. In A. Armando, R. Baldoni, & R. Focardi (Eds.), *Proceedings of the 1st Italian Conference on Cybersecurity (ITASEC17)* (pp. 86-95). CEUR.

Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. *ACM Computing Surveys,* 51(4), 85. https://doi.org/10.1145/3232676

Haidar, B., Chamoun, M., & Serhrouchni, A. (2017). A multilingual system for cyberbullying detection: Arabic content detection using machine learning. *Advances in Science, Technology and Engineering Systems Journal,* 2(6), 275-284. https://doi.org/10.25046/aj020634

Ho, Y., & Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access,* 8, 4806-4813. https://doi.org/10.1109/ACCESS.2019.2962617

Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A., & Bensaali, F. (2023). Deep transfer learning for automatic speech recognition: Towards better generalization. *Knowledge-Based Systems, 277*, 110851. https://doi.org/10.1016/j.knosys.2023.110851

Köffer, S., Riehle, D. M., Höhenberger, S., & Becker, J. (2018). *Discussing the value of automatic hate speech detection in online debates*. Paper presented at the Multikonferenz Wirtschaftsinformatik 2018 (MKWI 2018): Data Driven X-Turning Data in Value, Lüneburg, Germany.

Kshirsagar, R., Cukuvac, T., McKeown, K., & McGregor, S. (2018, October 31). Predictive embeddings for hate speech detection on Twitter. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 26-32). Association for Computational Linguistics.

Li, C., Zhan, G., & Li, Z. (2018, October 19-21). News text classification based on improved Bi-LSTM-CNN. *Proceedings of 9th International Conference on Information Technology in Medicine and Education* (pp. 890-893). IEEE.

Liu, L., Chen, J., Fieguth, P., Zhao, G., Chellappa, R., & Pietikäinen, M. (2019). From BoW to CNN: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1), 74-109. https://doi.org/10.1007/s11263-018-1125-z

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. *Social Science Computer Review,* 38(2), 128-146. https://doi.org/10.1177/0894439318791786

Mocanu, D., Rossi, L., Zhang, Q., Karsai, M., & Quattrociocchi, W. (2015). Collective attention in the age of (mis)information. *Computers in Human Behavior,* 51(Pt B), 1198-1204. https://doi.org/10.1016/j.chb.2015.01.024

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016, April 11-15). Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web* (pp. 145-153). ACM.

Patle, A., & Chouhan, D. S. (2013, January 23-25). SVM kernel functions for classification. *Proceedings of the 2013 International Conference on Advances in Technology and Engineering* (pp. 1-9). IEEE.

Pérez, J. M., Luque, F. M., Zayat, D., Kondratzky, M., Moro, A., Serrati, P. S., Zajac, J., Miguel, P., Debandi, N., Gravano, A., & Cotik, V. (2023). Assessing the impact of contextual information in hate speech detection. *IEEE Access,* 11, 30575-30590. https://doi.org/10.1109/ACCESS.2023.3258973

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence,* 48(12), 4730-4742. https://doi.org/10.1007/s10489-018-1242-y

Rahman, M. M., Balakrishnan, D., Murthy, D., Kutlu, M.,

& Lease, M. (2021). An information retrieval approach to building datasets for hate speech detection. *arXiv*, 2106.09775. https://doi.org/10.48550/arXiv.2106.09775

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine,* 47(1), 31-39. https://doi.org/10.17849/insm-47-01-31-39.1

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2018). Hateminers: Detecting hate speech against women. *arXiv,* 1812.06700. https://doi.org/10.48550/arXiv.1812.06700

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences,* 10, 1. https://doi.org/10.1186/s13673-019-0205-6

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July 28-August 2). The risk of racial bias in hate speech detection. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668-1678). Association for Computational Linguistics.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica,* 24(1), 12-18. https://doi.org/10.11613/BM.2014.003

Srba, I., Lenzini, G., Pikuliak, M., & Pecar, S. (2021). Addressing hate speech with data science: An overview from computer science perspective. In S. Wachs, B. Koch-Priewe, & A. Zick (Eds.), *Hate speech - Multidisziplinäre analysen und handlungsoptionen* (pp. 317-336). Springer.

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM -- A tutorial into long short-term memory recurrent neural networks. *arXiv,* 1909.09586. https://doi.org/10.48550/arXiv.1909.09586

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics,* 126(1), 157-179. https://doi.org/10.1007/s11192-020-03737-6

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *arXiv*, 1705.09899. https://doi.org/10.48550/arXiv.1705.09899

Waseem, Z., & Hovy, D. (2016, June 12-17). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In J. Andreas, E. Choi, & A. Lazaridou (Eds.), *Proceedings of the NAACL Student Research Workshop* (pp. 88-93). Association for Computational Linguistics.

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access,* 6, 13825-13835. https://doi.org/10.1109/ACCESS.2018.2806394

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018, June 1-6). Inducing a lexicon of abusive words – A feature-based approach. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 1046-1056). Association for Computational Linguistics.

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2018). Efficient KNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems,* 29(5), 1774-1785. https://doi.org/10.1109/TNNLS.2017.2673241

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web,* 10(5), 925-945. https://semantic-web-journal.net/content/hate-speech-detection-solved-problem-challenging-case-long-tail-twitter-1

JISTaP https://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 13(1): 70-84, 2025

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2025.13.1.5

How Do Open and Closed Access Journals Compare in Citations, Altmetrics, and Social Media Engagement for Pesticide Research?

Sivagami Murugappan
Department of Library and Information Science, Alagappa University, Karaikudi, India
E-mail: sivagami158028@gmail.com

Jeyshankar Ramalingam*
Department of Library and Information Science, Pondicherry University, Pondicherry, India
E-mail: jeyshankar71@gmail.com

ABSTRACT

This study analyzes the roles of altmetric and citation scores in open and closed-access pesticide research journals from 2013 to 2023, revealing key insights into impact metrics across publishing models. Traditional citations predominantly favor subscription-based journals, which account for 68.03% of the total citations. In contrast, green open-access journals excel in altmetric scores, primarily driven by social media engagement on platforms such as Mendeley, Twitter, and Facebook. Green open-access documents record the highest cumulative citations (13,143) and altmetric scores (5,768), suggesting greater online visibility and broader social reach. Statistical analysis shows no significant difference between altmetrics and citations, indicating that both metrics contribute complementary perspectives on research impact. Descriptive statistics highlight variations in citation patterns, with open-access journals showing a more concentrated distribution. Toxicology journals, where much pesticide research is published, are predominantly closed access, though citation patterns and altmetric attention vary by journal. Leading journals such as *Ecotoxicology and Environmental Safety* emphasize citations, while *Food and Chemical Toxicology* focuses on altmetrics, underscoring the dual approach to research visibility and impact in pesticide studies. These findings emphasize the evolving role of altmetrics in complementing traditional citations, especially for studies with high public and social relevance.

Keywords: pesticide exposure, toxicology, open access, closed access, subscription, altimetric score

Received: March 19, 2024 **Revised:** November 6, 2024
Accepted: November 17, 2024 **Published:** March 30, 2025

***Corresponding Author:** Jeyshankar Ramalingam
https://orcid.org/0000-0002-0186-0657
E-mail: jeyshankar71@gmail.com

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. Open Access articles are automatically archived in the Korea Institute of Science and Technology Information (KISTI)'s Open Access repository (AccessON). All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/). Under this license, the authors retain full ownership of their work, while permitting anyone to use, distribute, and reproduce the content in any medium, as long as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© 2025 [Sivagami Murugappan, Jeyshankar Ramalingam] This article is distributed under the Creative Commons Attribution License (CC BY), allowing unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. It is published by the Korea Institute of Science and Technology Information (KISTI).

1. INTRODUCTION

The term *pesticide* encompasses various compounds, including insecticides, fungicides, herbicides, rodenticides, molluscicides, nematicides, plant growth regulators, and various other pesticides (Aktar et al., 2009). Pesticides play a crucial role in advancing agricultural development, but their application can lead to both immediate and prolonged human health risks, as well as significant environmental challenges (Tudi et al., 2022). Therefore, it is essential to research pesticide application methods, exposure pathways, and their impacts on health and the environment. Merely engaging in research is insufficient to mitigate pesticide exposure; effective communication and dissemination strategies are also imperative. Researchers increasingly use blogs, wikis, Twitter, Facebook, LinkedIn, and other social media, along with open access (OA) journals and online repositories, to disseminate their findings (Cho, 2021a). Online reference management software and proficient networks are now significant to research processes that provide insights into the social influence of research beyond conventional citation methods (Azer & Azer, 2019).

Although citation counts are a long-established and reliable metric for measuring the academic influence of a research paper, they do not fully capture the research's wider impact beyond the academic community. Therefore, there is a need for tools that measure the wide-ranging societal impact of research (Cho, 2015; Cho, 2021b). Altmetrics, propelled by social media, offers an alternative by tracking views, downloads, readings, and mentions of an article across various platforms (Kırlangıç et al., 2022). Agriculture is quite essential to satisfy basic human needs and also has economic, social, cultural, and environmental dimensions. Disseminating pesticide research through social media maximizes its societal impact, influences policy decisions, raises public awareness, fosters interdisciplinary collaboration, and enhances communication strategies. This study analyzes the relationship between altmetric scores and citation scores and examines how OA status affects altmetric scores. This section presents the analysis of altmetric and citation scores to explore how open and closed-access (CA) journals impact the visibility and engagement of pesticide research. By evaluating these metrics together, the results highlight key patterns in research dissemination, accessibility, and scholarly influence within the field.

1.1. Overview of Publishing Models

OA is a scholarly publishing model in which research findings and information are made freely available to readers at no cost. In contrast, the subscription-based model requires readers to pay for access to research content. OA journals have different ways of publishing such as gold, hybrid, green, and bronze. In the golden route, the article publishing charge (APC) is paid by the author or their institution and the documents are available to the readers in full text. In the hybrid route, hybrid journals are subscription-based scholarly platforms where authors can pay an APC for OA to specific works, while others remain behind a paywall, requiring readers to pay for full-text access. In green access, authors self-archive their communication and it will be published publically after the embargo period. Publications in bronze OA are freely accessible on publishers' platforms, but do not include a Creative Commons Attribution License. As a result, they may be subject to modifications at any time, potentially limiting their accessibility and reusability (Cho, 2021b).

This study explores the following questions:

1.2. Research Questions

RQ1) What are the altmetric scores of highly cited articles on pesticide exposure?

RQ2) Do the highest altmetric attention scores correlate with the citation numbers of pesticide exposure articles?

RQ3) Does OA status affect the altmetric and citation scores of highly cited articles on pesticide exposure?

RQ4) Which sources exhibit significant concentration in pesticide research literature and impact?

2. LITERATURE REVIEW

Altmetrics capture research impact beyond traditional citations, reflecting public engagement through platforms like Twitter, Facebook, and Mendeley. Although existing literature has examined the relationship between altmetrics, OA, and citations in disciplines including social sciences and medicine, this relationship remains underexplored in pesticide research. This study addresses that gap by examining altmetric scores and citation performance in open and CA journals within pesticide literature, identifying trends, inconsistencies, and evolving dynamics between traditional and altmetric metrics.

2.1. Reviews Based on Highly Cited Articles

Cho (2021b) investigated altmetrics for highly cited

social science papers (638) from Scopus, evaluating correlations with citations and OA status. Altmetric scores of views, readers, blogs, Wikis, and Tweets linked with citations and OA were analyzed using Spearman correlation and Mann-Whitney U Tests. The consequences demonstrated a strong relationship between citations and readership, with OA papers showing higher altmetric engagement. The top 1,000 articles of *Nature* on social media websites and tools are examined by Ouchi et al. (2019) to evaluate and correlate the altmetric score and their bibliometric indicators. In that, 98.9% of articles were incorporated with social media. Mendeley, Wikipedia, and CiteULike are the most altmetric-scored social media sources, and a positive correlation between altmetric score and citation score was also found. Nisha et al. (2022) conducted a bibliometric analysis of the top 100 most cited studies in information science and library science, revealing that most of the top 100 cited research papers (80%) are in paid subscription journals, emphasizing the importance for libraries to maintain these subscriptions for student access to high-quality research. Highly cited articles were primarily from the USA, Canada, and UK, indicating the growing influence of social media in research dissemination. Saberi and Ekhtiyari (2019) examined highly cited documents in the library and information science (LIS) field as identified by Google Scholar, revealing a positive and significant correlation between capture metrics and Google Scholar citations. Similarly, Li and Thelwall (2012) conducted a comparative analysis of 1,397 genomics and genetics articles using Faculty of 1,000 and Mendeley data alongside traditional bibliometric indicators, identifying a significant correlation between citations and Journal Impact Factors.

Previous research on highly cited studies demonstrates a positive correlation between altmetric scores and traditional citations, with OA papers generally achieving greater altmetric presence and visibility.

2.2. Reviews Based on Open Access and Altmetric Score

The impact of OA on altmetrics for LIS research was determined by Cho (2021a), who found a prevalence of Mendeley bookmark readers (63%) and an unreliable comparison of altmetrics in OA papers. Particularly, views, blogs, and tweets display higher sensitivity in OA papers, while Mendeley readers shows a major correlation with citations. Mounce (2015) suggests that OA may offer an altmetric advantage due to increased visibility, noting that seven out of the top ten papers ranked by Altmetric.com in 2012 were OA. Mounce (2015) also points out that while the OA model is relatively well-established, altmetrics are still new and largely unexplored. Bernal (2013) reviews studies on OA and emerging metrics, highlighting that altmetrics provide added value for OA (Poplašen & Grgić, 2017) and that there is a correlation between page views, citations, and social media impact (Priem et al., 2012). It is obvious from the above literature that OA boosts altmetric impact by increasing visibility and engagement across platforms such as Mendeley and social media, strengthening research reach and influence.

2.3. Reviews Based on Pessimistic Correlation between Citation and Altmetric Score

Heydari et al. (2019) analyzed 1,077 highly cited surgery articles and found 62.74% of documents received altmetric scores. Positive correlations were observed between citations and Wikipedia, policy-making documents, CiteUlike, and citation count. A non-significant positive correlation was found with Mendeley readers. Overall, the weak and negative correlation between citation counts and altmetric attention scores indicates that altmetrics serve as a complementary measure rather than a replacement for traditional citations. Azer and Azer (2019) searched for top-cited articles on medical professionalism from the Web of Science database and identified that no correlation was found between citation and altmetric scores for publications before 2007, but for those published in 2007 and after, there was a significant correlation. The overall summary of the literature confirms that altmetrics generally show weak correlations with traditional citations, indicating that they complement rather than replace citations. Positive correlations were noted with platforms such as Wikipedia and policy documents, but with some variations depending on the publication year.

2.4. Reviews Based on Optimistic Correlation between Citation and Altmetric Score

Statistically significant associations were identified for metrics including Twitter, Facebook, research highlights, blogs, mainstream media, and forums. These findings underscore the influence of time in using altmetrics to rank articles and raise questions about the practical utility of altmetrics due to their inconsistent coverage (Thelwall et al., 2013). A study by Mohammadi and Thelwall (2014) examined the correlation between Mendeley readership counts and citations in social sciences and humanities disciplines, and also identified patterns of information flow between scientific fields using Mendeley data, showing

potential for capturing knowledge transfer across disciplines and providing early-stage impact evidence. Thelwall et al. (2023) conducted an evaluation of the UK Research Excellence Framework 2021 and found that altmetrics demonstrated a stronger correlation with research quality and field-normalized Scopus citation counts. Among the various altmetrics, Mendeley readers counts emerged as the most reliable indicator across the majority of assessment units. It was also found that altmetrics are generally stronger indicators of research quality in health and physical sciences and weaker in the arts and humanities.

Former studies show a clear correlation between the number of citations an article receives and its free online availability (Davis & Walters, 2011; Gargouri et al., 2010; Harnad & Brody, 2004; Kurtz et al., 2005; Lawrence, 2001). Very few studies have examined the connection between OA and altmetric scores. It is crucial to avoid applying the same criteria to older and newer articles. As Web 2.0 continues to evolve, the growing number of social media users has increased altmetric scores for earlier publications compared to more recent ones—a trend attributed to the expansion of communication channels rather than differences in publication quality (Adie, 2014). Moreover, scientific publications generally receive higher altmetric scores than those in other fields. Despite existing literature, significant gaps remain in understanding the correlation between altmetrics and citations, particularly in specialized fields such as pesticide research. Most studies focus on broad disciplines like social sciences and medicine, leaving the relationship between OA and altmetric scores within pesticide research underexplored. Inconsistencies across studies regarding the relationship between altmetrics and citations further highlight the need for more focused investigation.

This study addresses these gaps by examining the interplay among altmetrics, citations, and OA in pesticide research, providing new insights into the field. Previous research has shown strong associations between altmetrics and social media activity, with timing playing a crucial role in article rankings. Mendeley readership demonstrates a significant correlation with citation counts, and altmetrics serve as stronger indicators of research quality in health sciences than in the arts. However, few studies have explored the link between OA and altmetrics in specialized areas such as pesticide research. The following sections outline the study's materials, methods, and analytical approaches used to achieve these objectives.

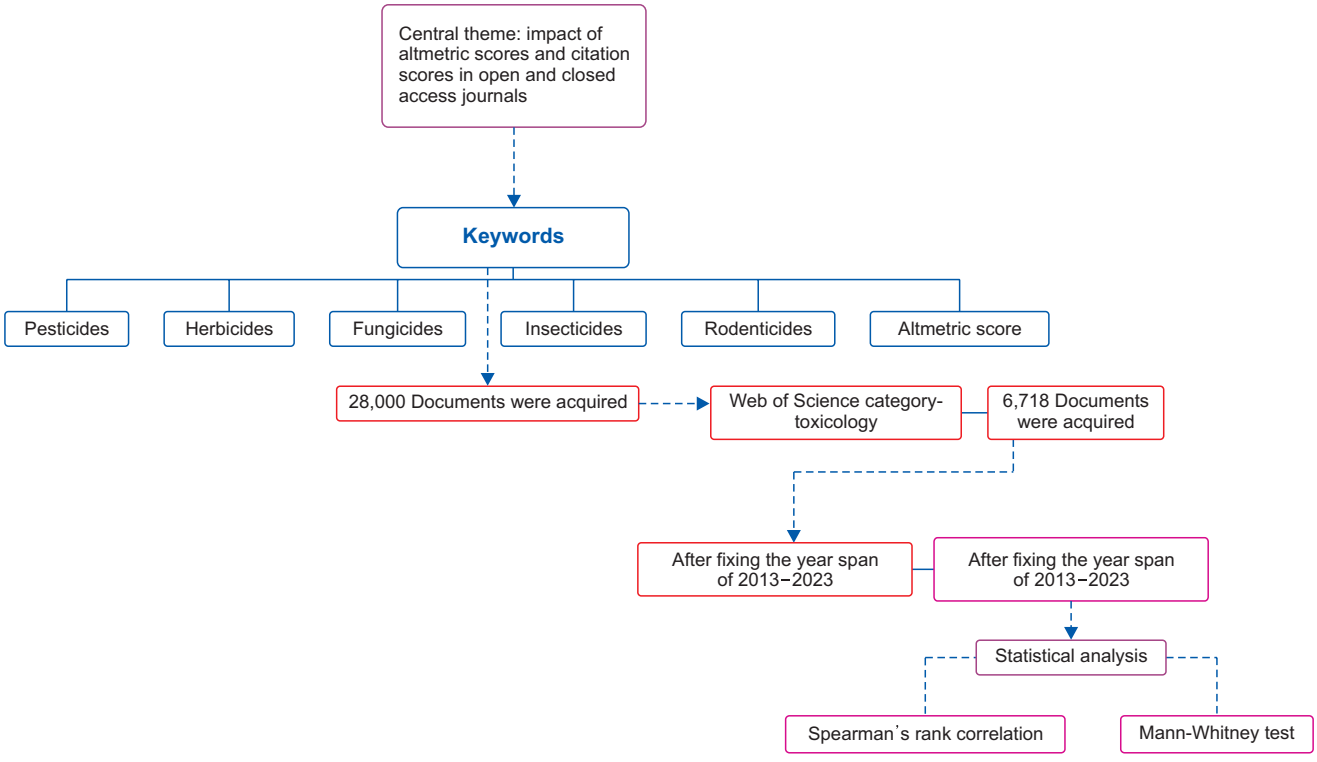


Fig. 1. Data structure of the study.

3. METHODOLOGY

3.1. Materials

To address the research questions, data were retrieved from the Web of Science database, maintained and published by Clarivate Analytics. Although Scopus and Google Scholar also offer citation tracking capabilities, it was essential to limit the search to the Web of Science due to its consistent updates and amendments. On December 20, 2023, data were collected using the keywords 'Pesticides,' 'Herbicides,' 'Fungicides,' 'Rodenticides,' 'Insecticides,' and 'Altmetric Score' in the basic search query. This initial query yielded 28,000 documents. To refine the results, the Web of Science category 'Toxicology' was applied, narrowing the dataset to 6,718 documents. Further filtering was performed by restricting the publication years to 2013-2023 to focus on recent developments in traditional citation scores and social media metrics within the pesticide research literature. Ultimately, 3,291 documents were selected for analysis (Fig. 1).

3.2. Methods

The citation and statistical analyses were conducted using Bibexcel software (2017 ver; Olle Persson, Leuven, Belgium), while the altmetric scores were obtained from the *Journal of Altmetrics* (https://www.journalofaltmetrics.org) by utilizing the DOIs of all 3,291 documents, resulting in altmetric scores for 1,961 documents. To determine the association between citation scores, altmetric scores, and OA status, Spearman's rank correlation analysis was performed using IBM SPSS Statistics 26 (IBM Corp., Armonk, NY, USA). Furthermore, the Mann-Whitney U Test was utilized to determine whether OA articles received a significantly higher number of citations and altmetric scores compared to CA articles (Fig. 2).

4. RESULTS

Integrating altmetric and citation scores enables researchers to gauge both societal impact and academic significance in pesticide research, providing a more nuanced understanding of broader implications and practical applications. This dual analysis enhances the formal demonstration of research significance.

Integrating altmetric and citation scores enables researchers to gauge both societal impact and academic significance in pesticide research, providing a more nuanced understanding of broader implications and practical applications. This dual analysis enhances the formal demonstration of research significance. Table 1 presents data encompassing 3,291 documents spanning the period of 2013-2023, contributed by 104 journals. Remarkably, 79.82% of the documents are articles, 11.39% are meeting abstracts, and 5.40% are reviews. Document types other than this account are less than 2% of the total publications.

4.1. Traditional Citation

A substantial portion of documents (68.03%) is published in subscription or CA journals, contrasting with 31.97% exclusively in OA journals. Within OA, prevalence is observed in green access (603 documents, 13,143 citations), followed by hybrid access (170 documents, 3,123 citations), gold access (158 documents, 2,488 citations), and bronze access (121 documents, 2,150 citations). Research focus is especially concentrated in subscription-based journals with green OA being the preferred choice

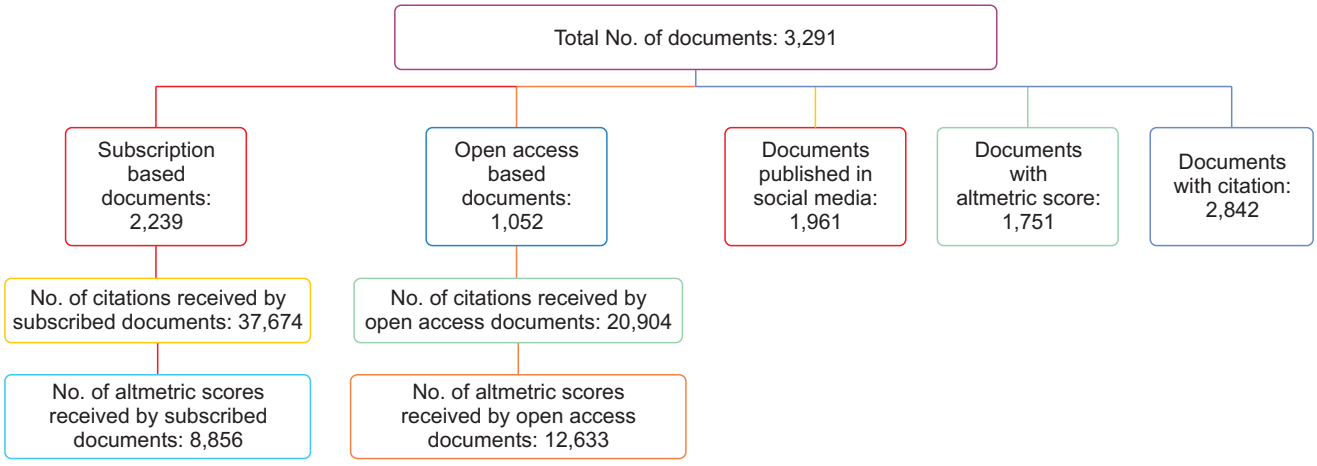


Fig. 2. Overview of collected data: Key information visualization.

 https://doi.org/10.1633/JISTaP.2025.13.1.5

Table 1. Overview of document types and characteristics (2013-2023)

Description	Result
Total no. of documents	3,291
Time span	2013-2023
Sources (journals, books, etc.)	104
Annual growth rate (%)	-18.1
Document average age	5.41
Document type	
Article	2,627 (79.82%)
Meeting abstract	375 (11.39%)
Review	178 (5.40%)
Letter	45 (1.36%)
Correction	29 (0.88%)
Editorial material	29 (0.88%)
News item	5 (0.15%)
Retraction	3 (0.09%)
Status	
Subscription based documents	2,239
Open access documents	1,052
Documents media published in social	1,961
Documents with altmetric scores	1,751
Documents with citation	2,842
No. of citations received by open access documents	20,904
No. of citations received by subscribed documents	37,674
No. of altmetric scores received by open access documents	12,633
No. of altmetric scores received by subscribed documents	8,856

for its absence of article publication charges, and the highest citation was received in the post-embargo period. The cumulative citation count across 3,291 documents stands at 58,578, with the majority of citations occurring in 2013, accounting for 11,049 citations. Publications in recent years are yet to be cited. Subscription journals led with 37,674 citations, while OA journals received 20,904 citations. Traditional citation scores indicate 2,842 cited documents, where most of the documents are articles and 449 meeting abstract documents are uncited.

4.2. Altmetric Score

Among the 3,291 documents, 1,961 have been integrated into social media, predominantly comprising articles and chapters. Of these, 1,751 documents have received altmetric scores, while 210 are pending evaluation. OA journals have garnered the highest altmetric scores, totaling 12,633 from 782 documents, with green access leading at 5,768 scores from 226 documents. Gold access follows with 3,599 scores (221 documents), hybrid access with 2,834 scores (195 documents), and bronze access with 432 scores (100 documents), reflecting fewer publications engaging with social media. Subscription-based journals received 8,856 altmetric scores from 1,179 documents.

Table 2 presents descriptive statistics for traditional citations received by pesticide research literature, categorizing documents into subscribed and OA. Subscribed documents show a publication range of 24 to 274, with a mean of 203.55 (standard deviation [SD]=66.21) representing a moderate level of variability around the average, while OA documents range from 25 to 155 publications, with a mean of 95.64 (SD=37.22) representing more concentrated distribution around the average. For subscribed documents citation, citations range from 45 to 8,125, with a mean of 3,424.91 (SD=2,331.53) indicating a wider variability in citation counts. OA citation documents have citations ranging from 53 to 3,122, with a mean of 1,900.36 (SD=925.67) designating a more concentrated distribution

Table 2. Descriptive statistics of traditional citation scores

Nature of access	Total no. of documents	No. of documents	Minimum	Maximum	Mean	Standard deviation
Subscribed documents	3,291	2,239	24	274	203.55	66.21
Open access documents	3,291	1,052	25	155	95.64	37.22
Subscribed documents citation	3,291	2,239	45	8,125	3,424.91	2,331.53
Open access citation	3,291	1,052	53	3,122	1,900.36	925.67

of citation values around the mean. These statistics reveal differences in citation patterns, indicating varying degrees of attention from researchers and reflecting diverse levels of impact, influence, or recognition within the academic community. In contrast, lower dispersion suggests a more consistent level of citation across the publications.

Table 3 outlines altmetric scores for a dataset of 1,961 documents, which has various dimensions of online attention. Mendeley readers mean scores (51.97, SD=71.448) display the highest number of times a research document has been saved or 'read' by users, followed by Dimensions citations (27.38, SD=43.660). The number of citation count of 22.92 was less than Mendeley readers and Dimensions citations, but higher than news, blog, Twitter, and Facebook mentions. While analyzing the SD counts, Facebook (89.324) obtains varying level of attention from researchers, followed by Twitter mention (87.214), Mendeley readers (71.448), and Dimensions citations (43.660). While citation is a significant metric for traditional scholarly impact, the altmetric scores, representing online attention and engagement, divulge that on average, documents may obtain more attention on platforms such as Facebook, Mendeley, Dimensions citations, and Twitter compared to traditional scholarly citations.

Levene's test was conducted to evaluate the assumption of equal variances between OA and CA journals. A *p*-value of 0.102, which exceeds the typical significance threshold of 0.05, indicates that there is no significant difference in variance between the two groups. The t-test was subsequently applied to assess whether a significant difference exists in mean citation counts between OA and CA journals. The negative t-test value indicates that the mean citation count for OA journals is lower than that of CA journals. However, despite both the Levene's test and t-test showing *p*-values below 0.05—implying statistical significance—the observed difference in mean citations did not translate into a significant variation in citation rates between OA and CA journals. The 95% confidence intervals provide a range within which the true difference in means likely falls, underscoring the importance of cautious interpretation (Table 4).

A Spearman's rank correlation analysis was conducted to examine the relationship between citations and altmetric scores across various sources, including news, blogs, Twitter, Facebook, Mendeley readers, Dimensions citations, and total citations (Table 5). The results showed that blog mentions had a weak positive correlation with traditional citations, while Dimensions citations, Mendeley readers, and Facebook mentions demonstrated strong correlations above 40%. Twitter mentions had the highest

Table 3. Descriptive statistics of altmetric scores

Altmetric scores	Total no. of documents	No. of altmetric scored documents	Minimum	Maximum	Mean	Standard deviation
News mentions	1,961	1,751	0	203	0.65	5.496
Blog mentions	1,961	1,751	0	54	0.19	1.580
Twitter mentions	1,961	1,751	0	3,083	7.93	87.214
Facebook mentions	1,961	1,751	0	3,196	3.79	89.324
Mendeley readers	1,961	1,751	0	1,336	51.97	71.448
Dimensions citations	1,961	1,751	0	821	27.38	43.660
Citations	1,961	1,751	0	708	22.92	36.299

Table 4. Difference in number of citations in open access and subscribed access articles - independent samples test

Total citations	Levene's test for equality of variances		t-test for equality of means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean difference	Standard error difference	95% Confidence interval of the difference	
								Lower	Upper
Equal variances assumed	2.68	0.102	-2.7	3,289	0.007	-3.044	1.127	-5.255	-0.834

 https://doi.org/10.1633/JISTaP.2025.13.1.5

Table 5. Correlation between altmetric scores and citation

		News mentions	Blog mentions	Twitter mentions	Facebook mentions	Mendeley readers	Dimensions citations	Citations
News mentions	Pearson correlation	1	0.794	0.696	0.039	0.355	0.332	0.314
	Sig. (2-tailed)		0.000	0.000	0.083	0.000	0.000	0.000
	n	1,961	1,961	1,961	1,961	1,961	1,961	1,961
Blog mentions	Pearson correlation	0.794	1	0.679	0.132	0.248	0.209	0.210
	Sig. (2-tailed)	0.000		0.000	0.000	0.000	0.000	0.000
	n	1,961	1,961	1,961	1,961	1,961	1,961	1,961
Twitter mentions	Pearson correlation	0.696	0.679	1	0.264	0.157	0.111	0.099
	Sig. (2-tailed)	0.000	0.000		0.000	0.000	0.000	0.000
	n	1,961	1,961	1,961	1,961	1,961	1,961	1,961
Facebook mentions	Pearson correlation	0.039	0.132	0.264	1	0.079	0.064	0.064
	Sig. (2-tailed)	0.083	0.000	0.000		0.000	0.005	0.004
	n	1,961	1,961	1,961	1,961	1,961	1,961	1,961
Mendeley readers	Pearson correlation	0.355	0.248	0.157	0.079	1	0.940	0.504
	Sig. (2-tailed)	0.000	0.000	0.000	0.000		0.000	0.000
	n	1,961	1,961	1,961	1,961	1,961	1,961	1,961
Dimensions citations	Pearson correlation	0.332	0.209	0.111	0.064	0.940	1	0.499
	Sig. (2-tailed)	0.000	0.000	0.000	0.005	0.000		0.000
	n	1,961	1,961	1,961	1,961	1,961	1,961	1,961
Citations	Pearson correlation	0.314	0.210	0.099	0.064	0.504	0.499	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.004	0.000	0.000	
	n	1,961	1,961	1,961	1,961	1,961	1,961	1,961

Table 6. Mann-Whitney U Test statistics for altmetric scores and citation

	News mentions	Blog mentions	Twitter mentions	Facebook mention	Mendeley readers	Dimensions citations	Citation
Mann-Whitney U	426,746.500	434,309.500	383,213.500	439,557.500	428,577.000	455,421.500	414,101.000
Z	-5.175	-4.746	-6.469	-3.132	-2.640	-0.454	-3.820
Asymp. Sig. (2-tailed)	0.000	0.000	0.000	0.002	0.008	0.650	0.000

correlation with citations at 99%, suggesting that researchers may bookmark these works for future reference.

New mentions were highly correlated with all social media sources except Facebook (0.03). Blog mentions showed weak correlations with most social media sources, except for strong correlations with news (79%) and Twitter (67%). Twitter had weak associations with Facebook, Mendeley, and Dimensions citations but was more closely linked with blogs (69%) and news (67%). Facebook exhibited minimal connections to news, blogs, and Twitter but had strong correlations with Mendeley and Dimensions citations. Mendeley readers were highly correlated with Dimensions citations (94%) but weakly correlated with other sources. Similarly, Dimensions citations had strong associations with Mendeley and Facebook but weaker correlations with other mentions.

The determination of statistically significant differences between altmetric scores and citations is verified

through the Mann-Whitney U Test, as depicted in Table 6. The results reveal significant differences found in altmetric scores and citations, except for Dimensions citations (0.650), which is evidenced by *p*-values less than the predictable significance level of 0.05. Negative Z-scores suggest that the altmetric scores are generally lower than the citations. The Mann-Whitney U Test results emphasize prominent divisions in the relationship between altmetric scores and citations across distinct categories.

Table 7 summarizes the differences in altmetrics and citations between CA (subscription-based) and OA journals across various categories. CA articles outperformed OA articles in news mentions, with a higher average rank (951.96 vs. 1,024.79) and total performance (1,122,356.50 vs. 801,384.50). A similar trend was observed for blog mentions, where CA articles also had a higher average rank (958.37 vs. 1,015.12) and total performance (1,129,919.50 vs. 793,821.50). However, OA articles performed better in Twitter mentions, with a higher sum rank despite a lower mean rank (915.03). For overall citations, OA articles had a higher average rank (1,040.96 vs. 941.23) but lower total performance (814,030.00 vs. 1,109,711.00), suggesting greater visibility and impact for OA publications. Lower values for both metrics generally reflect better performance, emphasizing the increased reach of OA articles within scholarly and online communities.

Table 8 provides a comprehensive assessment of publication metrics for diverse toxicology journals, enabling a thorough comparison of their performance in terms of visibility, impact, and accessibility. Notably, the top 10 journals exhibit significant contributions from the Netherlands and the USA, both recognized as leading exporters of agricultural products. Among these, the journal *Ecotoxicology and Environmental Safety*, based in the USA, emerges as the foremost contributor, marked by extensive

Table 7. Open access differences in altmetrics and citations

NOA of AS and citation		n	Mean rank	Sum of ranks
News mentions	Closed	1,179	951.96	1,122,356.50
	Open access	782	1,024.79	801,384.50
	Total	1,961		
Blog mentions	Closed	1,179	958.37	1,129,919.50
	Open access	782	1,015.12	793,821.50
	Total	1,961		
Twitter mentions	Closed	1,179	915.03	1,078,823.50
	Open access	782	1,080.46	844,917.50
	Total	1,961		
Facebook mentions	Closed	1,179	962.82	1,135,167.50
	Open access	782	1,008.41	788,573.50
	Total	1,961		
Mendeley readers	Closed	1,179	953.51	1,124,187.00
	Open access	782	1,022.45	799,554.00
	Total	1,961		
Dimensions citations	Closed	1,179	976.28	1,151,031.50
	Open access	782	988.12	772,709.50
	Total	1,961		
Citation	Closed	1,179	941.23	1,109,711.00
	Open access	782	1,040.96	814,030.00
	Total	1,961		

NOA, number of open access; AS, altmetric score.

 https://doi.org/10.1633/JISTaP.2025.13.1.5

Table 8. Top 10 highly cited journals and their impact through citation and altmetric scores

Journals (country)	Open access	Closed	Total doc	TC	Open access-doc	Closed- doc	Total doc	Total AS	Impact factor (H-Index)
Ecotoxicology and Environmental Safety (USA)	161 (2,648)	221 (6,558)	382	9,206	134 (531)	160 (355)	294	886	7.129 (161)
Environmental Toxicology and Chemistry (USA)	88 (1,634)	175 (3,152)	263	4,786	78 (627)	152 (1,194)	230	1821	4.218 (186)
Aquatic Toxicology (Netherlands)	44 (810)	99 (2,436)	143	3,246	35 (94)	67 (324)	102	418	4.964 (156)
Ecotoxicology (Netherlands)	33 (611)	135 (2,316)	168	2,927	26 (242)	110 (478)	136	720	2.935 (101)
Food And Chemical Toxicology (UK)	39 (673)	77 (1,766)	116	2,439	33 (781)	43 (1,763)	76	2,544	5.572 (192)
Environmental Toxicology and Pharmacology (Netherlands)	11 (234)	82 (1,684)	93	1,918	7 (104)	45 (574)	52	678	5.785 (97)
Bulletin of Environmental Contamination and Toxicology (USA)	18 (297)	148 (1,330)	166	1,627	14 (81)	102 (196)	116	277	2.807 (80)
Food Additives and Contaminants Part A-Chemistry Analysis Control Exposure & Risk Assessment (UK)	12 (128)	88 (1,408)	100	1,536	7 (9)	40 (66)	47	75	3.549 (57)
Toxicology Letters (Netherlands)	13 (144)	286 (1,378)	299	1,522	5 (18)	30 (640)	35	658	4.372 (158)
Clinical Toxicology (UK)	18 (177)	80 (181)	98	358	16 (145)	19 (90)	35	235	3.738 (103)

TC, total citation; AS, altmetric score.

citations and a notable H-Index (161). Despite its prominence, the journal predominantly publishes under CA, showcasing a robust Impact Factor of 7.129. However, it is observed that this journal has a limited focus on social media engagement. Conversely, the UK-based *Food and Chemical Toxicology* stands out with the highest altmetric score of 2,544, complemented by substantial H-Index values (192) and a commendable Impact Factor of 5.572. The journal's prominence on social media is remarkable, evident through a significant altmetric score, and its contributions, though predominantly in CA journals, receive the highest citations (1,766). An overarching observation is that these impactful works are predominantly featured in subscription-based journals, indicating widespread subscription among various institutions. This information is invaluable for researchers, institutions, and readers keen on understanding the influence and reach of these journals within the field of toxicology.

Table 9 unveils a curated list of the top 10 research articles concerning the impact of pesticides on human health. The most cited article, "Pesticides and human chronic diseases: Evidence, mechanisms, and perspectives," authored by S. Mostafalou and published in the journal *Toxicology and Applied Pharmacology* in 2013, has garnered the highest citation count of 708. Despite its citation success, the article has a relatively modest altmetric score of 251, suggesting untapped potential for greater social media visibility and impact are highlighted in bold letters. A pattern emerges where top articles, primarily published in the journal *Toxicology*, receive comparatively lower altmetric scores. This indicates an opportunity for these impactful research works to enhance their visibility and societal impact by leveraging social media platforms. Such efforts could contribute to increased awareness and engagement

Table 9. Top 10 highly cited articles and their altmetric scores

Rank	Title (type of access)	Source	First author	Year	AS	Citation
1	Pesticides and human chronic diseases: Evidences, mechanisms, and perspectives (closed access)	*Toxicology and Applied Pharmacology*	Sara Mostafalou	2013	251	708
2	Heavy metals and pesticides toxicity in agricultural soil and plants: Ecological risks and human health implications (open access)	*Toxics*	Ahmed Alengebawy	2021	245	419
3	Toxic effects of pesticide mixtures at a molecular level: Their relevance to human health (closed access)	*Toxicology*	Antonio F. Hernandez	2013	10	375
4	Pesticides: An update of human exposure and toxicity (closed access)	*Archives of Toxicology*	Sara Mostafalou	2017	64	370
5	Impact of glyphosate and glyphosate-based herbicides on the freshwater environment (closed access)	*Journal of Applied Toxicology*	Robert Annett	2014	38	319
6	Neurodevelopmental disorders and prenatal residential proximity to agricultural pesticides: The CHARGE study (open access)	*Environmental Health Perspectives*	Janie F. Shelton	2014	426	315
7	Persistent organochlorinated pesticides and mechanisms of their toxicity (closed access)	*Toxicology*	Ezra J. Mrema	2013	4	303
8	Effects of neonicotinoid pesticide exposure on human health: A systematic review (open access)	*Environmental Health Perspectives*	Andria M. Cimino	2017	187	276
9	Hazardous effects of chemical pesticides on human health-cancer and other associated disorders (open access)	*Environmental Toxicology and Pharmacology*	Akash Sabarwal	2018	63	263
10	Ethoxylated adjuvants of glyphosate-based herbicides are active principles of human cell toxicity (closed access)	*Toxicology*	Robin Mesnage	2013	128	252

Top 3 high altmetric score received articles, and their scores are as follows: Pesticides and human chronic diseases: Neuro developmental disorders and prenatal residential proximity to agricultural pesticides: The CHARGE study (open access, 426); Pesticides and human chronic diseases: Evidences, mechanisms, and perspectives (closed access, 251); Heavy metals and pesticides toxicity in agricultural soil and plants: Ecological risks and human health implications (open access, 245).
AS, altmetric score.

 https://doi.org/10.1633/JISTaP.2025.13.1.5

Table 10. Top 10 highest altmetric scored articles and their citations

Rank	Title	Source	First Author	Year	Citation	AS
1	Exposure to glyphosate-based herbicides and risk for non-Hodgkin lymphoma: A meta-analysis and supporting evidence	*Mutation Research/ Reviews in Mutation Research*	L. Zhang	2019	156	2,928
2	Refined assessment and perspectives on the cumulative risk resulting from the dietary exposure to pesticide residues in the Danish population	*Food and Chemical Toxicology*	Olaf Larsson Martin	2018	12	1,307
3	Cytotoxicity on human cells of Cry1Ab and Cry1Ac Bt insecticidal toxins alone or with a glyphosate-based herbicide	*Journal of Applied Toxicology*	Robin Mesnage	2013	40	659
4	Neurodevelopmental disorders and prenatal residential proximity to agricultural pesticides: The CHARGE study	*Environmental Health Perspectives*	Janie F. Shelton	2014	315	426
5	An acute exposure to glyphosate-based herbicide alters aromatase levels in testis and sperm nuclear quality	*Environmental Toxicology and Pharmacology*	E. Cassault-Meyer	2014	73	425
6	Occupational exposure to pesticides and consequences on male semen and fertility: A review	*Toxicology letters*	O. Mehrpour	2014	149	413
7	Mechanisms underlying the neurotoxicity induced by glyphosate-based herbicide in immature rat hippocampus: Involvement of glutamate excitotoxicity	*Toxicology*	Diane Cattani	2014	151	361
8	Pesticides and human chronic diseases: Evidences, mechanisms, and perspectives	*Toxicology and Applied Pharmacology*	Sara Mostafalou	2013	708	251
9	Heavy metals and pesticides toxicity in agricultural soil and plants: Ecological risks and human health implications	*Toxics*	Ahmed Alengebawy	2021	419	245
10	Carbamate insecticides target human melatonin receptors	*Chemical Research in Toxicology*	M. Popovska-Gorevski	2017	36	226

AS, altmetric score.

among both researchers and non-academic audiences. These findings offer valuable insights into the performance and characteristics of leading toxicology journals, as well as the potential for maximizing the societal impact of influential research articles on pesticide-related human health impacts.

Table 10 represents the top articles that received more altmetric scores. The article "Exposure to glyphosate-based herbicides and risk for non-Hodgkin lymphoma: A meta-analysis and supporting evidence," published in 2019, received the highest altmetric score of 2,928 with the least citation number of 156. This suggests that the article's amalgamation in social media has created more impact than in traditional citation within the short period. The article "Heavy metals and pesticides toxicity in agricultural soil and plants: Ecological risks and human health implications," which was published in the recent year 2021, received more citations and received a moderate altmetric score specifying the quality of work. It can be merged into social media to get more visibility. Most journals specializing in toxicology-related subject areas occupy top-ranking positions. The highest altmetric-scored articles are recently published in the years 2018 to 2023, with a mere number of citations showing the dominance of social media against traditional citation.

5. DISCUSSION

The altmetric scores of highly cited articles are generally moderate, indicating the need for greater incorporation of these articles across various social media platforms. Currently, most of the articles are primarily associated with Mendeley readers and Dimensions citations. However, if these articles were more widely shared on platforms such as Facebook, Twitter, and blogs, they could potentially achieve higher altmetric scores, increasing their visibility and amplifying their impact on society (Heydari et al., 2019). For instance, the most-cited article, "Pesticides and

human chronic diseases: Evidence, mechanisms, and perspectives" by S. Mostafalou, published in *Toxicology and Applied Pharmacology* in 2013, has received 708 citations. However, its altmetric score is 251, with approximately 200 of these derived from Mendeley and Dimensions. These results suggest that while altmetrics, particularly on platforms like Mendeley, Dimensions, and Twitter, can indicate academic influence, there are clear distinctions between the engagement measured by altmetrics and traditional academic citations. Altmetric scores may reflect public interest and early-stage impact, but they are not a direct replacement for citation metrics in evaluating research impact. The findings emphasize that researchers need to strategically engage with a broader range of social media platforms to enhance both societal and academic visibility.

The analysis reveals nuanced differences between OA and CA journals in terms of citations and altmetric impact. While CA journals outperform OA journals in citations, OA journals gain greater visibility and impact on platforms like Twitter. The nature of OA content may facilitate its dissemination, particularly in online and social media spaces, even if this does not always result in higher citation counts. These findings highlight the complexity of measuring scholarly impact and suggest that different access models may influence distinct dimensions of academic visibility and engagement.

The study also notes that the highest altmetric articles are from toxicology journals and those published between 2018 and 2023, demonstrating the increasing influence of social media over traditional citation methods. Out of 3,291 total documents, only 60% (1,691) are featured on social media, with 71% of OA documents having non-zero altmetric scores, compared to 88.3% for CA documents. Further analysis reveals that OA articles generally perform better in altmetric scores, but show no significant difference in citation rates compared to CA articles. Moreover, the study finds that 30% of highly cited documents are not featured on social media, while the remaining 70% with higher altmetric scores are prominently present on platforms such as Twitter, Mendeley, and Dimensions. The comparison of altmetric and citation scores between OA and CA journals, with a p-value of 0.102, indicates no significant evidence to reject the assumption of equal variances. However, the negative t-test value of less than 0.05 suggests a meaningful difference between open and CA journals, consistent with previous research (Nisha et al., 2022). In inference, integrating traditional and non-traditional metrics provides a more comprehensive view of an article's impact. As new methods and trends evolve, they increasingly shape scholarly communication and must be monitored to effectively assess research significance.

The findings indicate that while there is no significant difference in citation rates between OA and CA journals, there is a noticeable distinction in their altmetric performance. OA articles tend to have a greater impact on social media platforms, potentially leading to higher visibility and wider dissemination among researchers and the general public. However, CA journals often exhibit better performance in traditional citation metrics. This highlights the emerging trend of social media outpacing traditional citations in terms of immediate visibility, particularly for newer research. Researchers and institutions may need to strategically utilize social media to maximize the reach and societal impact of their work.

6. CONCLUSION

The findings demonstrate a clear distinction between the performance of OA and CA journals in terms of citations and altmetric scores. While OA articles tend to perform better on social media platforms, CA journals still dominate in traditional citation metrics, especially within institutional and academic circles. For toxicology journals, the analysis highlights the opportunity for leveraging social media to enhance the visibility and societal impact of impactful research articles, especially those related to human health and pesticide exposure. Researchers and journals should consider utilizing altmetric data to expand their audience reach and foster greater engagement across diverse platforms. This integrated approach, combining both traditional citation analysis and altmetric evaluation, offers a more holistic understanding of the influence and visibility of research in the scholarly community and beyond.

7. LIMITATIONS

The data were exclusively retrieved from the Web of Science database for the period 2013-2023 and were specifically filtered using the Web of Science category 'Toxicology.' These limitations should be considered in future research, and subsequent studies may expand the scope by including additional databases and categories to provide a more comprehensive analysis.

 https://doi.org/10.1633/JISTaP.2025.13.1.5

FUNDING

A special thanks to our Alagappa University for providing an AURF Research Fellowship of Rs. 5,000 per month for the period of 2022-2024, which aided in facilitating the execution of this study.

ACKNOWLEDGMENTS

The author M. Sivagami extend heartfelt gratitude to N. N. Ghazzali and Aasif Ahmad Mir (research scholars), Department of Library and Information Science at Pondicherry University, for invaluable assistance in gathering altmetric scores.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Adie, E. (2014). *Attention! A study of open access vs non-open access articles*. https://www.altmetric.com/blog/attention-a-study-of-open-access-vs-non-open-access-articles/

Aktar, M. W., Sengupta, D., & Chowdhury, A. (2009). Impact of pesticides use in agriculture: Their benefits and hazards. *Interdisciplinary Toxicology*, 2(1), 1-12. https://doi.org/10.2478/v10102-009-0001-7

Azer, S. A., & Azer, S. (2019). Top-cited articles in medical professionalism: A bibliometric analysis versus altmetric scores. *British Medical Journal Open*, 9(7), e029433. https://doi.org/10.1136/bmjopen-2019-029433

Bernal, I. (2013). Open access and the changing landscape of research impact indicators: New roles for repositories. *Publications*, 1(2), 56-77.

Cho, J. (2015). A study about scholarly impact measurement through altmetrics. *Journal of Korean Library and Information Science Society*, 46(1), 65-81. https://doi.org/10.16981/kliss.46.1.201503.65

Cho, J. (2021a). Altmetrics analysis of highly cited academic papers in the field of library and information science. *Scientometrics*, 126(9), 7623-7635. https://doi.org/10.1007/s11192-021-04084-w

Cho, J. (2021b). Altmetrics of highly cited research papers in social science. *Serials Review*, 47(1), 17-27. https://doi.org/10.1080/00987913.2021.1882652

Davis, P. M., & Walters, W. H. (2011). The impact of free access to the scientific literature: A review of recent research. *Journal of the Medical Library Association*, 99(3), 208-217. https://doi.org/10.3163/1536-5050.99.3.008

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS One*, 5(10), e13636. https://doi.org/10.1371/journal.pone.0013636

Harnad, S., & Brody, T. (2004). Comparing the impact of open access (OA) vs. non-oa articles in the same journals. *D-Lib Magazine*, 10(6). http://www.dlib.org/dlib/june04/harnad/06harnad.html

Heydari, S., Shekofteh, M., & Kazerani, M. (2019). Relationship between altmetrics and citations: A study on the highly cited research papers. *DESIDOC Journal of Library & Information Technology*, 39(4), 169-174. https://www.proquest.com/openview/3139001d7c9d8253fbf31757e8f39b1d/1?pq-origsite=gscholar&cbl=2028807

Kırlangıç, M. M., Sahin, E., Sahin, M. E., Kütük, S., Halıcı, B. N. A., & Ak, M. (2022). Evaluation of altmetric analysis scores of the top 100 articles on polycystic ovary syndrome published in the last 10 years. *Cureus*, 14(12), e32903. https://doi.org/10.7759/cureus.32903

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. S. (2005). The effect of use and access on citations. *Information Processing & Management*, 41(6), 1395-1402. https://doi.org/10.1016/j.ipm.2005.03.010

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*, 411(6837), 521. https://doi.org/10.1038/35079151

Li, X., & Thelwall, M. (2012). *F1000, Mendeley and traditional bibliometric indicators*. Paper presented at 17th International Conference on Science and Technology Indicators, Montréal, QC, Canada.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, 65(8), 1627-1638. https://doi.org/10.1002/asi.23071

Mounce, R. (2015). Dark research: Information content in many modern research papers is not easily discoverable online. *PeerJ Preprints*, e773v771. https://doi.org/10.7287/peerj.preprints.773v1

Nisha, F., Kumar, S., Awasthi, S., & Tripathi, M. (2022). Top 100 cited papers in information and library science: A brief analysis. *Journal of Scientometric Research*, 11(3), 371-383. https://doi.org/10.5530/jscires.11.3.40

Ouchi, A., Saberi, M. K., Ansari, N., Hashempour, L., & Isfandyari-Moghaddam, A. (2019). Do altmetrics correlate with

citations? A study based on the 1,000 most-cited articles. *Information Discovery and Delivery*, 47(4), 192-202. https://doi.org/10.1108/IDD-07-2019-0050

Poplašen, L. M., & Grgić, I. H. (2017). Altmetric and bibliometric scores: Does open access matter? *Qualitative and Quantitative Methods in Libraries,* 5(2), 451-460. http://www.qqml.net/index.php/qqml/article/view/331

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv*, 1203.4745v1201. https://doi.org/10.48550/arXiv.1203.4745

Saberi, M. K., & Ekhtiyari, F. (2019). Usage, captures, mentions, social media and citations of LIS highly cited papers: An altmetrics study. *Performance Measurement and Metrics*, 20(1), 37-47. https://doi.org/10.1108/PMM-10-2018-0025

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS One*, 8(5), e64841. https://doi.org/10.1371/journal.pone.0064841

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2023). Do altmetric scores reflect article quality? Evidence from the UK Research Excellence Framework 2021. *Journal of the Association for Information Science and Technology*, 74(5), 582-593. https://doi.org/10.1002/asi.24751

Tudi, M., Li, H., Li, H., Wang, L., Lyu, J., Yang, L., Tong, S., Yu, Q. J., Ruan, H. D., Atabila, A., Phung, D. T., Sadler, R., & Connell, D. (2022). Exposure routes and health risks associated with pesticide application. *Toxics*, 10(6), 335. https://doi.org/10.3390/toxics10060335

JISTaP https://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 13(1): 85-98, 2025

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2025.13.1.6

Enhancing QA System Evaluation: An In-Depth Analysis of Metrics and Model-Specific Behaviors

Heesop Kim*
Department of Library Information Science, Kyungpook National University, Daegu, Korea
National Library of Korea, Seoul, Korea
E-mail: heesop@knu.ac.kr; heesopkim@korea.kr

Aluko Ademola
Department of Library Information Science, Kyungpook National University, Daegu, Korea
E-mail: alukojoshua@knu.ac.kr

ABSTRACT

The purpose of this study is to examine how evaluation metrics influence the perception and performance of question answering (QA) systems, particularly focusing on their effectiveness in QA tasks. We compare four different models: BERT, BioBERT, Bio-ClinicalBERT, and RoBERTa, utilizing ten EPIC-QA questions to assess each model's answer extraction performance. The analysis employs both semantic and lexical metrics. The outcomes reveal clear model-specific behaviors: Bio-ClinicalBERT initially identified irrelevant phrases before focusing on relevant information, whereas BERT and BioBERT continually converge on similar answers, exhibiting a high degree of similarity. RoBERTa, on the other hand, demonstrates effective use of long-range dependencies in text. Semantic metrics outperform lexical metrics, with BERTScore attaining the maximum accuracy (0.97), highlighting the significance of semantic evaluation. Our findings indicate that the choice of evaluation metrics significantly influences the perceived efficacy of models, suggesting that semantic metrics offer more nuanced and insightful assessments of QA system performance. This study contributes to the field of natural language processing and machine learning by providing guidelines for selecting evaluation metrics that align with the strengths and weaknesses of various QA approaches.

Keywords: question answering systems, evaluation metrics, natural language processing, transformer models, BERT

Received: September 7, 2024 **Revised:** November 12, 2024
Accepted: November 17, 2024 **Published:** March 30, 2025

***Corresponding Author:** Heesop Kim
https://orcid.org/0000-0002-4642-4445
E-mail: heesopkim@korea.kr

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. Open Access articles are automatically archived in the Korea Institute of Science and Technology Information (KISTI)'s Open Access repository (AccessON). All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/). Under this license, the authors retain full ownership of their work, while permitting anyone to use, distribute, and reproduce the content in any medium, as long as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© 2025 [Heesop Kim, Aluko Ademola] This article is distributed under the Creative Commons Attribution License (CC BY), allowing unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. It is published by the Korea Institute of Science and Technology Information (KISTI).

1. INTRODUCTION

The goal of natural language processing (NLP), a field within computer science and artificial intelligence, is to bridge the gap between humans and machines by enabling the latter to comprehend, interpret, and produce human language. In today's data-driven world, NLP is essential due to the massive volumes of natural language data generated through chat logs, emails, and social media platforms. NLP is widely used in many fields, such as machine translation, speech recognition, named entity recognition, sentiment analysis, and question answering (QA). GPT-4, a potent language model, outperforms its predecessor, GPT-3, according to Baktash and Dawodi (2023). While GPT-4 offers numerous potential applications, it also presents drawbacks, including significant data and computational requirements, as well as ethical considerations.

1.1. Question Answering System

The goal of QA is to build automated systems that can understand natural language questions and provide accurate and concise answers. QA systems work by analyzing a user's question and then searching for relevant information in a given knowledge base, such as a database, a collection of documents, or the Internet. Unanue et al. (2021) analyze a QA system using the BERTScore metric. When fine-tuning a strong baseline, experiments conducted over four diverse language pairs yielded improvements of up to 0.58 percentage points (3.28%) in the Bilingual Evaluation Understudy (BLEU) score and up to 0.76 percentage points (0.98%) in the BERTScore. Three approaches were proposed for generating soft predictions that allow the network to remain completely differentiable end-to-end.

In order to assess QA systems, this research examines how several metrics—such as Accuracy, Precision, Recall, F1 Score, Metric for Evaluation of Translation with Explicit ORdering (METEOR), and BERTScore—affect system performance. The study aims to improve the accuracy and reliability of QA systems, specifically deep learning models such as bidirectional encoder representations from transformers (BERT), biomedical language representation model (BioBERT), Bio-ClinicalBERT, and RoBERTa. This analysis is conducted using the epidemic question answering (EPIC-QA) text analysis conference (TAC) 2020 dataset, a benchmark dataset commonly used for evaluating the performance of these models, as demonstrated by their success on benchmarks like SQuAD (see more details at https://bionlp.nlm.nih.gov/epic_qa/).

1.2. Objectives

The primary aim of this study is to explore how various evaluation metrics impact the assessment of QA systems. Specifically, it seeks to address the following objectives: to identify how different evaluation metrics affect the perception of the strengths and weaknesses of QA systems, and to determine the most effective metrics for evaluating QA systems by comparing lexical and semantic metrics.

2. LITERATURE REVIEW

2.1. Performance Assessment Metrics for QA Systems

Metrics for automatic assessment are vital instruments for evaluating QA system performance. These metrics aid in the quantitative measurement of how well these systems perform in terms of delivering efficient and accurate responses when compared to the ground truth. Raza et al. (2022) used two datasets in this work: a reference-standard dataset built using the CORD-19 and LitCOVID initiatives, and a gold-standard dataset prepared by the experts from a public health domain, to assess their QA system, CoQUAD, using both lexical and semantic answer similarities. According to their research, the retrieval pipeline performs better when the top@k is increased, resulting in 83.6% recall in the top 20 out of the eight metrics used to assess the QA system. A variety of criteria are used to assess the efficacy of QA systems to precisely gauge performance. Research has explored numerous methods for improving these assessments. For example, Deutsch et al. (2021) presented a QA evaluation during the 2020 TAC for COVID-19 questions using normalized discount cumulative gain (NDCG) and normalized discount novel score (NDNS). The efficacy of NDNS was demonstrated by the HLTRI and h2oloo teams, achieving Exact scores of 0.371 and 0.368, respectively.

2.2. Evaluation Metrics

Different components of performance in QA are measured using lexical metrics and semantic similarity. While semantic similarities evaluate underlying meaning and context, lexical metrics concentrate on text features like word overlap. Both are beneficial and require specific methods. Bowman and Dahl (2021) discuss the present benchmarking procedures for natural language understanding, contending that these practices are problematic and do not provide an adequate evaluation of linguistic aptitude. They put forth four essential requirements for benchmarks to be responsible and effective: (1) Adver-

 https://doi.org/10.1633/JISTaP.2025.13.1.6

sarial Filtering: The conventional method used to gather benchmark datasets is questioned since it might not fully reflect language comprehension; (2) Direct Addressing of Criteria: The main parameters used to gauge language comprehension should be specifically addressed by benchmarks; (3) Expert Integration: The integration of expert effort into crowdsourced data collection is crucial for improving the quality of benchmarks; and (4) Institutional Model: Establishing a clear institutional model to address bias in metrics is identified as an open research question.

2.2.1. Lexical Metrics

Lexical metrics are based on surface-level features of text, such as word frequency, length, and grammatical structures. North et al. (2023), in their review of computational methods for lexical complexity prediction in English texts, emphasized the application of machine learning techniques and various features to identify complex terms and suggest simpler alternatives. They also highlighted the use of lexical complexity prediction in readability and text simplification. These metrics help assess how well the language used in the prediction response compares to the reference summary in terms of similarities, relevance, and appropriateness.

Consider the following scenario to answer the question "What are the symptoms of COVID-19?": The prediction answer correctly names fever, coughing, and dyspnea as the three primary COVID-19 symptoms. However, it does not address other common symptoms such as muscle aches, fatigue, loss of taste or smell, and sore throat. While the response includes some significant symptoms, it needs to be expanded for increased accuracy. The reference summary provides basic information but lacks crucial details. Lexical metrics offer quantitative insights into how well the prediction answer conveys the essential details and significance of the reference summary, indicating areas where the prediction model needs improvement. These measurements mainly emphasize factors like word overlap, term frequency, and structural similarity. This study will use several criteria to assess the lexical quality of QA systems, including ROUGE, F1 Score, exact match (EM), and Accuracy. By using these lexical metrics, we aim to assess how well QA systems capture the essential information and language patterns of the reference answers.

2.2.1.1. ROUGE.

ROUGE metrics evaluate the quality of QA systems by comparing generated responses against human-written ones and calculating the overlap between predicted and ground truth answers. Yang et al. (2018) assessed a QA system designed to respond to factoid queries using ROUGE metrics. Their research indicated that their modifications of ROUGE measurements had a stronger alignment with human judgment compared to the original ROUGE-L. For QA, the most widely utilized ROUGE metrics are ROUGE-1, ROUGE-2, and ROUGE-L.

2.2.1.1.1. ROUGE-1.

ROUGE-1 measures the overlap of unigrams (single words) between a system-generated answer and a set of reference answers.

$$\text{ROUGE-1}=\frac{Number\ of\ overlapping\ unigrams}{Total\ number\ of\ unigrams\ in\ reference\ answers} \cdots\cdots (1)$$

For example, given the reference answer: "*The cat sat on the mat*" and the system-generated answer: "*The cat on the rug*," the overlapping unigrams are "*The*," "*cat*," and "*on*." The total unigrams in the reference answer are 6 (as "*sat*," "*the*," and "*mat*" are not present in the system-generated answer). Therefore ROUGE-1=$\frac{3}{6}$=0.5. Through the evaluation of word overlap, these metrics assist in quantifying the degree to which the prediction accurately captures the important information from the reference answer.

2.2.1.1.2. ROUGE 2.

ROUGE-2 measures bigram overlaps by counting the number of consecutive word pairs that match between the reference and prediction. This aids in assessing the coherence and fluency of responses produced by the system.

$$\text{ROUGE-2}=\frac{Number\ of\ overlapping\ bigrams}{Total\ number\ of\ bigrams\ in\ reference\ answers} \cdots\cdots (2)$$

For example, given the reference answer: "*The cat sat on the mat*" and the system-generated answer: "*The cat on the rug*," the overlapping bigrams are "*The cat*" and "*on the*." The total number of bigrams in the reference answers is 5 (as "*The cat*," "*cat sat*," "*sat on*," "*on the*," and "*the mat*"). Therefore, ROUGE-2=$\frac{2}{5}$=0.4. Because word order is important for preserving textual meaning and flow in QA systems, ROUGE-2 offers a more rigorous evaluation than ROUGE-1.

2.2.1.1.3. ROUGE L.

ROUGE-L calculates the longest common subsequence (LCS) between a collection of reference answer(s) and an answer generated by the system. In contrast to ROUGE-1 and ROUGE-2, which assess word or bigram overlap, ROUGE-L accounts for sentence structure and word

order by capturing the longest sequence of words that appear in both the system-generated and reference answers in the same order.

$$\text{ROUGE-L}=\frac{LCS(X,Y)}{m} \quad \cdots\cdots (3)$$

Where: X is the system-generated answer, Y is the set of reference answers, and LCS(X,Y) is the length of the LCS between X and Y. m is the length of the reference answer (or the average length if multiple references are used).

For example, given the reference answer: "*The cat sat on the mat*" and the system-generated answer: "*The cat on the rug*," and LCS: "*The cat sat on the*," length of reference (m) is 6. Therefore, ROUGE-L=$\frac{5}{6}$=0.8.

2.2.1.2. F1 Score.

Question-answering systems are frequently assessed using the F1 Score. This is determined by examining the precision and recall of the system's predictions. D'Hoffschmidt et al. (2020) created the French QA dataset FQuAD using 25,000 question-answer pairs from Wikipedia. A baseline model achieved 82.1 EM and 92.2 F1 Scores on the test set. The F1 Score is calculated as follows:

$$F1=2\times\frac{Precision\times Recall}{Precision+Recall} \quad \cdots\cdots (4)$$

For example, given a reference answer with 10 correct answers and a system-generated answer with 8 answers (6 of which are correct), the Precision=$\frac{6}{8}$=0.75, and Recall=$\frac{6}{10}$=0.60. Therefore, F1=2×$\frac{0.75\times0.60}{0.75+0.60}$=0.66.

2.2.1.3. Accuracy.

Accuracy measures the proportion of correct answers generated by a QA system out of all the answers it produced.

$$\text{Accuracy}=\frac{Number\ of\ correct\ answers\ generated\ by\ the\ system}{Total\ number\ of\ answers\ generated\ by\ the\ system} \cdots (5)$$

For example, given a reference set of 10 questions, each with a single correct answer, and a system-generated set of 10 answers, 8 of which are correct, the Accuracy=$\frac{8}{10}$=0.8.

2.2.1.4. Exact Match.

EM is the percentage of questions where the system-generated response precisely matches one of the reference solutions, character by character. Wang et al. (2022) created a framework to assess the reading comprehension skills of machines. The assessment is easy to grasp and takes into account several factors. Tests demonstrate the approach's significant improvement, showing increases in EM and F1 Scores on MRC tasks of up to 11.22% and 8.71%, respectively.

$$\text{Exact match (EM)}=\frac{Number\ of\ questions\ with\ exact\ matches}{Total\ number\ of\ questions} \cdots (6)$$

For example, given the question *"What is the capital of France?,"* with the reference answer *[Paris]* and the system-generated answer *[Paris]*, the EM score is 1 (perfect match). In another example, given the question *"What is the tallest mountain in the world?,"* with the reference answer *[Mount Everest]* and the system-generated answer *[Alps]*, the EM score is 0 (no EM).

2.2.2. Semantic Answer Similarities Metrics

Metrics for semantic similarity assess how closely the meanings of the system-generated responses match those of the reference answers. Herbold (2024) devised the STSScore method, which uses an optimized model to directly estimate semantic similarity between texts. They discovered that STSScore performs better than conventional techniques based on embeddings or subsequence overlap. A common approach in COVID-19 QA systems is to find the cosine similarity between a question and potential answers, both represented as vector embeddings.

For example, given the question *"What are the symptoms of COVID-19?,"* the system-generated prediction is *"COVID-19 symptoms include fever, cough, and difficulty breathing,"* while the reference summary is *"COVID-19 is a virus that can cause severe respiratory illness."*

Using a pre-trained language model, these statements can be represented as vectors. It is then possible to determine the cosine similarity between each answer vector and the question vector. Amur et al. (2023)'s review of short-text semantic similarity demonstrates the potential of deep learning. They surveyed datasets and noted shortcomings in conventional techniques. The response with the highest cosine similarity is considered to be more semantically comparable to the query. Metrics such as BERTScore, Bi-Encoder, Cross-Encoder, and METEOR are important indicators of semantic similarity. However, BERTScore and METEOR are the main focus of our investigation.

2.2.2.1. BERTScore.

BERTScore is a metric used for evaluating the similarity between two pieces of text, typically in the context of NLP tasks such as QA. It is based on the BERT model, a

 https://doi.org/10.1633/JISTaP.2025.13.1.6

neural network model used for language representation. Zhang et al. (2020) propose BERTScore as an automatic evaluation metric for text generation, similar to common metrics, by calculating a similarity rating for each token in the candidate sentence relative to every word in the reference sentence. They assessed the outputs of 363 machine translation and image captioning systems using BERTScore, which showed a stronger correlation with human evaluation and outperformed existing evaluation metrics in selecting the best model. Chen and Eger (2023) introduce new evaluation metrics based on natural language inference (NLI) to address vulnerabilities in BERT-based metrics that are susceptible to adversarial attacks. These NLI-based metrics demonstrate greater robustness and outperform existing summarization metrics, though they still fall below the state-of-the-art for machine translation. Combining NLI metrics with existing ones enhances both adversarial robustness and performance on standard benchmarks. Precision and recall are combined to compute an F1 measure in this context.

For a reference x and candidate $\hat{x}$, the Recall, Precision, and F1 Scores are defined as follows:

Recall: This measures the proportion of relevant tokens in the reference sentence that are also present in the candidate sentence. A high recall score indicates that the generated sentence covers most of the semantic content of the reference sentence. The recall is calculated as follows:

$$R_{BERT} = \frac{1}{|x|}\sum_{x_j \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \quad (7)$$

Precision: This measures the proportion of relevant tokens in the candidate sentence that are also present in the reference sentence. A high precision score indicates that most of the generated tokens are semantically similar to the reference tokens. The calculation is shown below:

$$p_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_j \in x} x_i^T \hat{x}_j \quad (8)$$

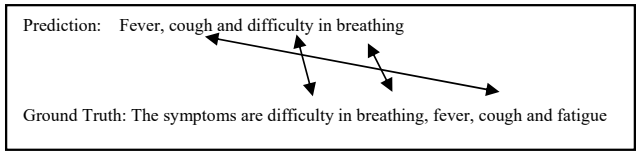
F1 measure: This is the harmonic mean of precision and recall. It provides a balanced measure of the accuracy of the generated sentence, taking into account both the semantic similarity of the generated tokens and how well the generated sentence covers the content of the reference sentence. See below on how it is calculated:

$$F_{BERT} = 2\frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}} \quad (9)$$

2.2.2.2. METEOR.

METEOR is widely used in NLP for evaluating language translation and QA tasks by comparing generated responses to reference answers, scoring based on their similarity and overall quality. Soni and Roberts (2022) proposed a method for generating clinical question variations based on EHRs. Results showed METEOR outperformed BLEU in paraphrase detection, with METEOR scoring 21.27 and BLEU 13.25. METEOR uses lexical matching, stemming, and synonymy to assess overlap between answers, considering word order, phrasing, and grammar. It adjusts its formula to account for tokenization and alignment with reference answers. METEOR balances precision and recall, incorporating factors like EM, stems, synonyms, and paraphrases. This approach allows METEOR to recognize different ways of expressing the same idea, making it more robust in handling language variations. One of METEOR's key advantages is its strong correlation with human judgment, as it evaluates both the accuracy and completeness of a translation. Its flexibility across languages and domains, along with its customizability, makes it valuable for various NLP applications beyond machine translation, such as text summarization and paraphrase detection.

The formula for METEOR is as follows:



The information above shows there are lexical similarities between the prediction and the ground truth and the demonstrates that METEOR also captures semantic similarity when evaluating QA. Computing the overall METEOR score combines the F-score computed from precision and recall.

$$M = F_{mean}(1-p) \quad (10)$$

2.3. Transformers Model (Baseline Approach)

NLP has evolved tremendously with the introduction of the Transformer model by Vaswani et al. (2017), especially for QA tasks. This deep learning model functions through an encoder-decoder architecture. The encoder

processes the input sequence (such as a question) to create hidden representations that capture the context and meaning of each token. The decoder then uses these representations to generate the output sequence, which includes the answer. Our research employs four distinct Transformer model iterations: BERT, BioBERT, Bio-ClinicalBERT, and RoBERTa. Each model was selected based on its specific strengths and performance in handling biomedical text and QA tasks (Fig. 1).

2.3.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a transformer-based approach created specifically for language understanding applications. Unlike the original Transformer model, BERT is trained bidirectionally, meaning it simultaneously considers context from both the left and right sides of a word. This bidirectional training allows BERT to more effectively grasp complex relationships and nuanced meanings of words. Alzubi et al. (2023) recommended DistilBERT, a faster and lighter version of BERT, for COVID-19 queries. After being fine-tuned using the SQuAD 1.1 dataset, DistilBERT outperformed earlier models, achieving an F1 Score of 87.3 and an EM score of 80.6.

2.3.2. Biomedical Language Representation Model (BioBERT)

BioBERT, based on BERT, is a specialized model for the biomedical domain, pre-trained on biomedical texts and fine-tuned for tasks including QA, named entity recognition, and relation extraction, outperforming other models. An investigation by Campos-Mercade et al. (2021) assessed various language models for QA related to COVID-19, revealing that BioBERT outperformed other models, particularly in tasks involving the filtering of sentences based on their relevance. In this research, we utilize the BioBERT model developed by the Data Mining and Information Systems Laboratory at Korea University, specifically designed for NLP tasks in the biomedical domain.

2.3.3. Bio-ClinicalBERT

Bio-ClinicalBERT, a BERT-based model pre-trained on biomedical literature, performs exceptionally well in QA challenges such as the BioASQ Challenge. This study improved Bio-ClinicalBERT's performance for specific tasks on the EPIC dataset by experimenting with differ-

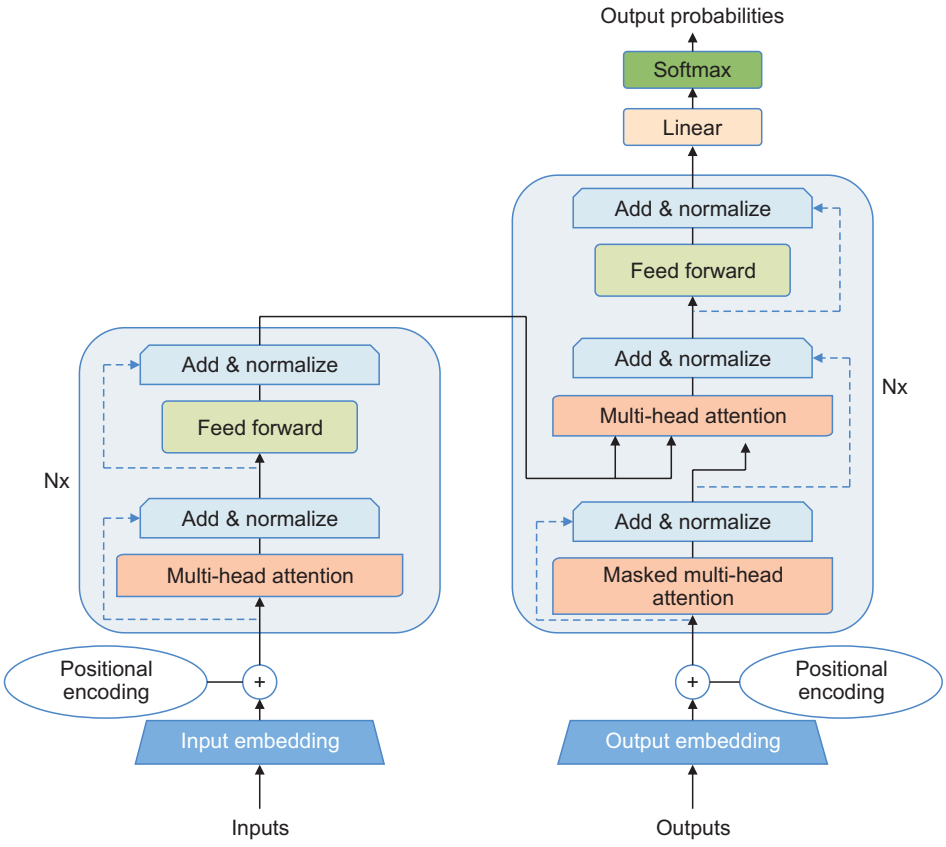


Fig. 1. The transformer model.

 https://doi.org/10.1633/JISTaP.2025.13.1.6

ent embeddings. Alsentzer et al. (2019) demonstrated that employing a model customized for the medical domain and developing embeddings with ClinicalBERT, a variation of Bio-ClinicalBERT, yields better results than non-domain-specific embeddings in three common clinical NLP tasks. In this study, the "emilyalsentzer/Bio_ClinicalBERT" model was employed to extract answers from the provided context.

2.3.4. RoBERTa

RoBERTa is a pre-trained language model built on the Transformer architecture, similar to BERT but with several significant enhancements to the training process. These improvements include dynamic masking and the elimination of the next sentence prediction task, both of which lead to stronger language comprehension. RoBERTa has been trained on a large corpus of text data to gain a thorough understanding of linguistic structures and patterns. Möller et al. (2020) significantly improved performance by fine-tuning a pre-trained RoBERTa model using the SQuAD and COVID-QA datasets, a QA dataset consisting of 2,019 question/answer pairs annotated by volunteer biomedical experts on scientific articles related to COVID-19. In this project, the RoBERTa model will be used to extract answers, leveraging its strong language understanding capabilities through a two-stage QA process (pre-training and fine-tuning).

3. RESEARCH DESIGN

3.1. Data Collections

This study uses the EPIC-QA dataset to assess the performance of various QA models in a specific domain. EPIC-QA focuses on ad-hoc questions related to diseases, with a specific emphasis on COVID-19, aligning with our research goal of analyzing QA models in a biomedical context. However, we acknowledge that the EPIC-QA dataset may not fully capture the entire range of QA system requirements. EPIC-QA delves deeper than generic QA datasets like SQuAD or MovieQA, which typically address factual information retrieval. The dataset is divided into two categories: Expert Level, designed to address questions relevant to professionals in the field, and Consumer Level, focused on answering questions posed by the general public. This study will concentrate on the Consumer QA subset, which pertains to the general public's inquiries about COVID-19 (Fig. 2).

```
[
  {                                    # First question in this collection
   "question_id": "EQ001",             # Globally unique question identifier
   "question": "what is the origin of COVID-19",  # Text of the question
   "query": "coronavirus origin",      # Search engine query associated with this question
   "background": "seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal
       ↪ source, and first transmission into humans"
  },
  {...},                               # Second question in this collection
  ...
]
```

Fig. 2. The structure of epidemic question answering.

3.2. Dataset Analysis

As part of the TAC, Goodwin et al. (2020) from U.S. National Library of Medicine launched the EPIC-QA track. The goal of this track was to further the study of automated question-answering systems, particularly those that deal with COVID-19. Creating systems that could modify their responses for various user groups, including specialists and the general public, was the main goal. To train and assess their systems, participants were given access to a variety of datasets, such as scholarly publications and easier-to-access information sources. The dataset analysis involves the following key sources: *Ask-Science*, which consists of 263 JSON files, each representing a post made between December 1, 2019, and October 29, 2020, in the /r/AskScience community on Reddit. These posts, tagged by community moderators, relate to COVID-19, health, medicine, or biology, and include one or more COVID-19 keywords. *CCNS-TREC* contains 114,645 JSON files, each corresponding to an HTML-parsed website included in the CommonCrawl News Subset used by the TREC Misinformation Track. These files contain one or more COVID-19 keywords. *CHQA* comprises 2,739 JSON files, corresponding to an October 9, 2020 HTML-parsed snapshot of web pages affiliated with the NIH, CDC, or WHO that contain COVID-19 keywords. To ensure a balanced and representative dataset for our ten questions, we sampled a subset of documents from each dataset. The final dataset used for analysis consisted of ten documents, including three from Ask-Science, three from CCNS-TREC, and four from CHQA.

3.3. Dataset Structure

The EPIC-QA dataset comprises two separate collections: answer extraction and end-to-end tasks. The dataset is well-structured and easily accessible for research, as each document in these collections is organized as a separate JSON file.

3.4. Structure of Context, Question, and Answer in QA

A QA system's goal is to produce a response to a specified inquiry based on a written passage. Contextualization is important because it establishes the framework for both the query and the response. This context may include historical background, relevant data, and other details that help the reader fully understand the issue at hand. Raza et al. (2022) used a COVID-19 QA dataset to illustrate an extractive QA system. This system takes a query as input and generates an answer by extracting relevant information from a given context. For example, the question "How long after you get exposed to the COVID-19 virus is a test able to detect?" is addressed with the following context: "The timing of when a COVID-19 test can detect the virus after exposure depends on several factors, such as the type of test used, the viral load in the body, and the individual's immune response. Generally, it can take between 2 to 14 days after exposure for a COVID-19 test to detect the virus in the body." The system generates the answer: "It can take between 2 to 14 days after exposure to COVID-19 for a test to detect the virus in the body."

3.5. Procedures

The QA system will be implemented using selected QA approaches, designed to dynamically generate answers to user queries by extracting information from a dataset, rather than relying on pre-stored answers in a database (Fig. 3).

3.5.1. Step 1: Data Loading

The process begins by loading relevant documents from the dataset that corresponds to the user's query. This is accomplished using a data loading library, such as PyTorch's DataLoader, which efficiently handles the dataset and ensures that only relevant documents are passed into the QA system.

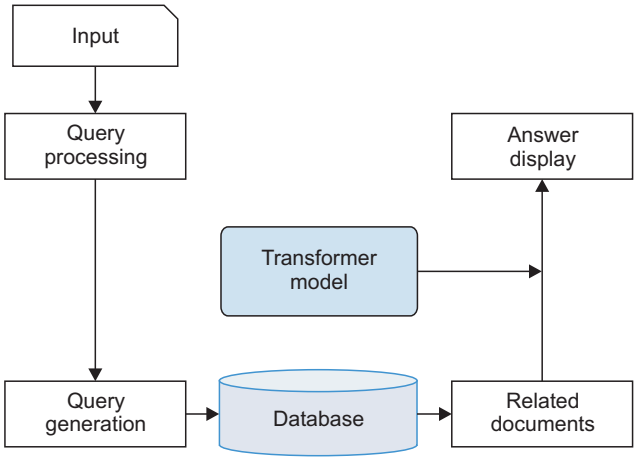


Fig. 3. Question answering pipeline architecture.

3.5.2. Step 2: Data Retrieval

Data retrieval is a crucial component of the QA process, essential for extracting relevant information from a corpus. Efficient and successful information retrieval is increasingly important in the era of big data, where vast amounts of data are generated daily. Several retrieval strategies are employed. Lexical Search (BM25) is a traditional approach based on keyword matching and term frequency. Semantic Search, on the other hand, uses more advanced techniques such as Dense Passage Retriever, Bi-Encoder, and Cross-Encoder, which leverage deep learning to understand the context and meaning behind words. Given the nature of this study and the structure of the EPIC-QA dataset, it will be necessary to integrate the retriever model with the reader model to accurately extract relevant answers.

3.5.3. Step 3: Answer Extraction (Reader Model)

The reader model is the last stage of the QA process. Its job is to comprehend the text and extract relevant information in order to provide a response. To respond, the reader divides the question into different portions, searches for essential words and phrases, and compares them to pertinent areas in the dataset. Responses are sent back to users once the reader determines the correct answer.

3.5.4. Step 4: Problem Formulation

The problem investigated in this research can be characterized as follows: The input question sequence (q) is represented as q=[q0, q1, q2, …, qn], where q0 through qn represent the individual tokens in the input question. The input context or passage sequence (p) is represented as p=[p0, p1, p2, …, pn], where p0 through pn represent the individual tokens in the context or passage from which the answer is to be extracted. The similarity function (S), denoted as S(q, p), measures the relevance of each token in the passage to the question. This function takes into account the contextual information of the tokens and is learned during the training process. Softmax is a normalization function that converts the relevance scores produced by S(q, p) into probabilities. Finally, argmax is a function that selects the token in the passage with the highest probability as the answer. These components work together in the QA system to identify and extract the most

 https://doi.org/10.1633/JISTaP.2025.13.1.6

relevant answer from the context provided.

3.6. Dataset Preprocessing

A thorough preprocessing workflow was applied to the dataset to ensure the efficacy of our QA model. The aim of this preprocessing was to standardize and purify the data, preparing it for training and assessment. During the data processing phase, several procedures were followed to prepare the text for analysis. First, the data was cleaned by removing HTML tags and special characters. Tokenization and lemmatization were then applied to standardize and reduce dimensionality. Stopwords were removed to emphasize more informative keywords. The dataset was divided into training, validation, and test sets to provide a fair and diverse assessment of the model. Finally, text sequences were encoded using BERT embeddings and padded to a constant sequence length to optimize batch processing during model training.

3.7. Experimental Setup

The experimental setup was designed to leverage state-of-the-art machine learning algorithms and hardware acceleration to achieve high accuracy in QA tasks. The desktop computer used was a high-performance device equipped with a robust CPU, GPU, and Coral USB accelerator, providing all the necessary hardware for efficient model training and inference. Python was used to implement TensorFlow 2.4.0 in Google Colab. The pre-trained models BERT, BioBERT, Bio-ClinicalBERT, and RoBERTa were selected for their capacity in quality assurance. The model's performance was evaluated using metrics including ROUGE, Accuracy, EM, and F1 Score.

3.7.1. Experimental Procedure

We conducted a comparative analysis of four pre-trained language models from the Transformers library—BERT, BioBERT, Bio-ClinicalBERT, and RoBERTa—targeted at QA tasks across three distinct domains: general, biomedical, and clinical.

3.7.2. Dataset and Preprocessing

We used a text analysis dataset specifically selected for the biomedical field to train these models. Preprocessing techniques were applied to the dataset to ensure the data was relevant and properly formatted for the QA tasks. These techniques included filtering out unnecessary information and applying sentence segmentation.

3.7.3. Model Fine-Tuning

Each model was fine-tuned on its respective training dataset using the AdamW optimizer. The fine-tuning process was configured with the following hyperparameters: a learning rate of 2e-5, a batch size of 32, and 5 epochs. To prevent overfitting, we implemented early stopping by monitoring the loss function, and training was halted if the loss did not improve for three consecutive epochs.

3.7.4. Evaluation Metrics

The performance of the models was evaluated using a combination of metrics that assessed accuracy, fluency, and semantic similarity. The F1 Score and Accuracy metrics measured the model's ability to correctly identify the correct answers, providing insights into precision and recall. The ROUGE-L Score was used to evaluate the overlap between the generated answers and the reference answers, focusing on the LCS to assess the quality of the generated text. The METEOR Score assessed the fluency and semantic coherence of the generated answers, taking into account synonyms, stemming, and word order. Lastly, BERTScore measured the contextual and syntactic similarity between the generated and reference answers, leveraging pre-trained BERT embeddings to evaluate the depth of understanding.

3.8. List of Questions

This section presents sample questions that showcase the QA system's ability to handle various COVID-19 inquiries. Although the number of questions in this experimental study is relatively small, they cover a range of topics, from infection rates to prevention.

1. Who can contract COVID-19?
2. Why are COVID-19 infection rates so low in Africa?
3. How do we prevent contracting COVID-19?
4. Is COVID-19 transmissible through mosquitos?
5. What makes some viruses seasonal?
6. When exactly do people during the illness start producing antibodies for COVID-19?
7. Which symptoms are commonly association with COVID-19?
8. Where should I go if I suspect I have symptoms?
9. Are children less susceptible to COVID-19 compared to adults?
10. What is the impact of COVID-19 on mental health?

4. RESULTS AND DISCUSSION

We used ten questions from the EPIC-QA dataset to assess four pre-trained language models—BERT, BioBERT, Bio-ClinicalBERT, and RoBERTa—due to our dataset's domain-specific nature. Metrics appropriate for biomedical writing, such as the F1 Score, ROUGE, METEOR, and BERTScore, evaluated how well they handled language peculiar to the field. Notably, all models produced EM and accuracy metrics of zero, which is why they were excluded from the displayed results. Pal (2023) introduced CLIFT, a new benchmark for clinical domain QA, and discovered that while deep learning models exhibit strong performance on initial test sets, their performance may deteriorate when exposed to distributional changes.

This outcome emphasizes the need for more reliable models and assessment metrics in the clinical domain. These results suggest that these metrics were not properly aligned with the significance of addressing biomedical questions in this context.

4.1. Results of F1 Score

BERT and BioBERT achieved moderate F1 Scores of 67%, demonstrating their ability to answer both general and biomedical questions, highlighting their considerable QA capabilities. RoBERTa, with an F1 Score of 33%, may require additional fine-tuning, while Bio-ClinicalBERT, scoring 48%, shows specialization in clinical contexts. The F1 Score analysis offers a comprehensive understanding of each model's relative strengths and weaknesses in biomedical QA (Table 1).

4.2. Results of ROUGE

BERT and BioBERT exhibit better response alignment with the ground truth, as evidenced by their respective ROUGE scores of 67%, indicating significant n-gram overlap and accuracy. Bio-ClinicalBERT and RoBERTa achieve 65% ROUGE scores, closely pursuing high alignment while having slightly different vocabulary and detail than BERT and BioBERT. In terms of obtaining significant factual components, BERT and BioBERT perform better than Bio-ClinicalBERT and RoBERTa, which show similar but considerably lower accuracy (Table 2).

4.3. Results of METEOR

The METEOR scores reveal that RoBERTa leads with 85%, indicating it generates the most fluent and coherent answers closely aligned with the ground truth. BioBERT follows at 82%, showing strong semantic similarity and relevance. Bio-ClinicalBERT scores 81%, reflecting robust domain understanding, while BERT, at 79%, slightly trails in conveying the intended meaning. Overall, these scores underscore each model's capability in producing meaningful and coherent responses, with RoBERTa excelling in fluency (Table 3).

4.4. Results of BERTScore

The BERTScore values show that BioBERT and Bio-ClinicalBERT achieved the highest scores (97%), demonstrating exceptional meaning and nuance matching with reference responses. BERT follows closely at 95%, indicating good semantic correctness. RoBERTa performs well with a score of 86%, though with some minor variations

Table 1. The results of F1 Score

Model name	Result (%)
BERT	67
BioBERT	67
Bio-ClinicalBERT	48
RoBERTa	33

Table 2. The results of ROUGE

Model name	Result (%)
BERT	67
BioBERT	67
Bio-ClinicalBERT	65
RoBERTa	65

Table 3. The results of METEOR

Model name	Result (%)
BERT	79
BioBERT	82
Bio-ClinicalBERT	81
RoBERTa	85

Table 4. The results of BERTScore

Model name	Result (%)
BERT	95
BioBERT	97
Bio-ClinicalBERT	97
RoBERTa	86

 https://doi.org/10.1633/JISTaP.2025.13.1.6

in finer aspects. Overall, BioBERT and Bio-ClinicalBERT excel at preserving semantic content, while BERT and RoBERTa demonstrate strong performance in capturing relevant information. The strong performance of RoBERTa in METEOR scores reflects its general effectiveness in NLP tasks, as discussed in various studies on transformer-based models. The exclusion of EM and Accuracy due to low scores is an important observation and may warrant further investigation into why these metrics performed poorly in this specific context (Table 4).

4.5. Maximum Scores of Each Metric

We evaluated QA models using BERTScore, METEOR, ROUGE, F1 Score, EM, and Accuracy. EM and Accuracy were excluded due to low scores. BERTScore achieved the highest at 97% with BioBERT and Bio-ClinicalBERT. METEOR reached 85% with RoBERTa, while F1 Score and ROUGE were 67% with BERT and BioBERT. These results highlight the relative strengths of each model across different evaluation metrics and provide a clearer picture of their effectiveness in answering questions (Fig. 4).

4.6. Evaluation Metrics and Answer Quality

In this section, we analyze the relationship between evaluation metrics and the quality of answers produced by the QA systems, illustrated below by one of the examples generated by the QA model and metric used in this study (Fig. 5).

4.7. Overall Performance of the QA Models

The performance variance among the four QA models highlights their unique strengths and weaknesses in predicting answers. This section answers Research Question (RQ) 1. BERT and BioBERT demonstrated exceptional capabilities in identifying relevant phrases. Bio-ClinicalBERT's reliance on contextual dependencies can lead to initial detection of irrelevant words, while RoBERTa's ability to handle long sequences and extract information from large contexts enhances its QA performance.

Overall, these models represent significant advancements in language model capabilities for QA tasks, enabling automated systems to capture context, understand queries, and provide more accurate and contextually aware answers across various applications. Zaib et al. (2021) propose a framework that integrates conversation history into a neural machine comprehension system using BERT. Results show that the proposed framework performs competitively with state-of-the-art models. However, they also highlight the negative impact of using the entire conversation history, as it introduces noise and unnecessary information, reducing model performance.

This analysis aligns well with current research trends in QA systems, as seen in the provided search results. The comparison of different models' strengths and weaknesses reflects the ongoing efforts to improve QA performance across various domains. The inclusion of Zaib et al. (2021)'s findings on conversation history integration adds valuable context to the challenges in developing more advanced QA systems.

4.8. Evaluation Metrics of the QA System

This study assessed the performance of BERT, BioBERT, Bio-ClinicalBERT, and RoBERTa models in predicting answers. The evaluation was conducted using lexical metrics such as F1 Score, EM, ROUGE, and Accuracy, as well as semantic metrics like METEOR and BERTScore. According to Turchin et al. (2023), biomedical BERT models—such as BioBERT and ClinicalBERT—perform better at recognizing complex medical concepts in medical documentation than standard BERT models. However, the ideal model selection may vary based on the specific task.

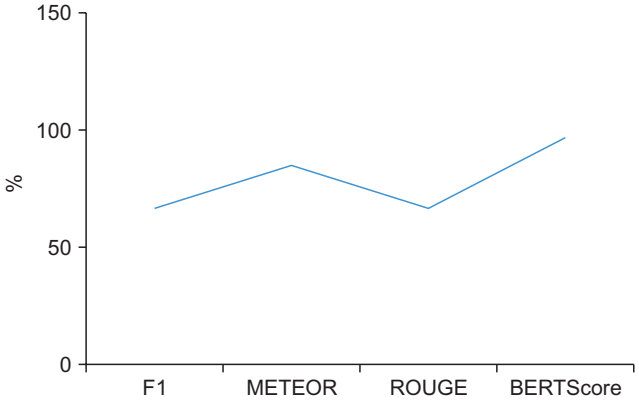


Fig. 4. Maximum scores of each metric.

```
from rouge import Rouge

def evaluate_rouge(reference_answers, predicted_answers):
    rouge = Rouge()
    scores = rouge.get_scores(predicted_answers, reference_answers)

    rouge_1_f1 = scores[0]['rouge-1']['f']
    rouge_2_f1 = scores[0]['rouge-2']['f']
    rouge_l_f1 = scores[0]['rouge-l']['f']

    return rouge_1_f1, rouge_2_f1, rouge_l_f1

# Example usage
reference_answers = ["children appear to be less susceptible"]
predicted_answers = ["Children are found to be less susceptible to COVID-19 compared to adults"]

rouge_1_f1, rouge_2_f1, rouge_l_f1 = evaluate_rouge(reference_answers, predicted_answers)

print("ROUGE-1 F1 score:", rouge_1_f1)
print("ROUGE-2 F1 score:", rouge_2_f1)
print("ROUGE-L F1 score:", rouge_l_f1)
```

```
ROUGE-1 F1 score: 0.49999999531250006
ROUGE-2 F1 score: 0.3749999953703125
ROUGE-L F1 score: 0.49999999531250006
```

Fig. 5. Evaluation metrics and answer quality.

Semantic measures were found to perform better than lexical metrics, addressing RQ2. BERTScore was notable for obtaining the highest score of 0.97 for the BioBERT and Bio-ClinicalBERT models. This result suggests that BERTScore, as a semantic metric, performed better than the other metrics included in this study when evaluating the quality of the responses. METEOR proved to be a robust metric for QA model evaluation, scoring 0.85 and capturing semantic nuances such as paraphrases, as well as EM. Even though BERTScore sometimes outperformed METEOR, the latter remains reliable. The QA system's accuracy and fluency are demonstrated by the ROUGE and F1 Scores of 0.67, which show a solid balance between precision and recall.

On the other hand, EM and Accuracy scores were low as they could not capture semantic equivalency. Models may perform poorly if their phrasing deviates from reference responses, even if the response is accurate and meaningful. To address these issues, the study employed more flexible and sophisticated assessment metrics, such as F1 Score, METEOR, and ROUGE. These metrics provide a comprehensive evaluation of the model's performance in QA tasks, considering token-level matching and measuring the precision and semantic quality of the generated answers.

The comparison of lexical and semantic metrics, as well as the focus on BERTScore and METEOR, reflects the ongoing efforts to improve the evaluation of QA models in the biomedical domain.

4.8.1. Implications

The varying performance of the QA models emphasizes the importance of selecting the appropriate model for specific applications. For tasks requiring precise phrase identification, BERT or BioBERT may be more suitable, while RoBERTa's ability to handle long and complex text could be advantageous in other contexts. Bio-ClinicalBERT's performance highlights the need for ongoing research to improve the robustness and accuracy of QA models. Addressing these shortcomings could lead to significant advancements in the field. The implications of these findings extend beyond academic study and into real-world applications. RoBERTa's ability to handle complex inputs could significantly enhance automated customer service by improving the model's capacity to process extended and more varied customer queries. Similarly, BioBERT's ability to accurately recognize medical terminology from patient records establishes it as a valuable asset in clinical decision support systems, potentially improving diagnostic accuracy and minimizing human error. As language models evolve, ongoing research to address the deficiencies of models like Bio-ClinicalBERT will be essential for enhancing QA systems across various industries.

4.8.2. Evaluation Metrics of the QA System

This study highlights a key issue: the difference between lexical and semantic metrics. Lexical metrics such as EM and Accuracy sometimes fail to account for semantic equivalence, penalizing models for minor variations in wording despite preserving meaning. This is particularly challenging in the biomedical domain, where nomenclature can differ substantially while representing the same context. As a result, models may receive low EM scores even when producing semantically accurate responses, underscoring the limitations of rigid lexical assessment in specialized tasks. Semantic metrics, such as BERTScore and METEOR, provide more flexibility by considering paraphrases. BERTScore showed exceptional performance for biomedical models (BioBERT, Bio-ClinicalBERT), effectively capturing semantic quality with a score of 0.97. METEOR demonstrated reliability for both paraphrases and precise matches, achieving a score of 0.85. These metrics offer a more nuanced assessment of QA models, particularly in complex domains like biomedical text, where understanding the intent and meaning of the question is crucial. While lexical measures are still useful for gauging fluency and accuracy, semantic metrics like METEOR and BERTScore are essential for evaluating the deeper quality of model responses, especially in specialized fields such as biomedicine. The results of this work support recent developments in the literature, such as Turchin et al. (2023), which highlight the importance of semantic equivalence over simple token matching and call for more sophisticated evaluation techniques.

5. CONCLUSION

In this study, we tested the effectiveness of various machine learning evaluation criteria to evaluate QA systems. The study by Bahak et al. (2023) evaluates ChatGPT's performance as a QA system in comparison to other models. It finds that while ChatGPT, based on GPT-3.5 Turbo, performs well on straightforward factual queries, its reliability decreases when dealing with more complex queries and when contextual information is lacking.

To determine their effect on assessing model quality, we selected and applied a variety of measures, including

 https://doi.org/10.1633/JISTaP.2025.13.1.6

lexical metrics (F1 Score, ROUGE, EM Accuracy), semantic and lexical metrics (METEOR), and semantic metrics (BERTScore). Our results showed that BERTScore is the best-performing metric for capturing similarity between prediction and ground-truth, based on contextual and grammatical nuances. METEOR also demonstrated strong performance, highlighting the importance of incorporating both semantic and lexical elements in QA assessment. F1 Score and ROUGE showed similar performance and provided valuable insights into the quality of the answers.

These measures are useful options for assessing QA systems as they cover important aspects of answer accuracy. EM and accuracy measurements did not perform well because they only consider word-level matching and ignore structure and contextual knowledge. This demonstrated their limitations by imposing penalties for even minor phrase modifications.

Since the efficacy of these metrics varies with different datasets and circumstances, alternative or supplementary evaluation approaches that consider context are necessary. Achieving accurate evaluation requires choosing metrics that align with the objectives and features of the QA system.

5.1. Limitations

While our study provides valuable insights into the performance of QA models, it is important to acknowledge the limitations of the research. Our results are not as broadly applicable as they could be due to the small sample size of ten questions taken from the EPIC-QA dataset. Expanding the dataset to include questions from different domains might be a consideration for future research to ensure more reliable results. Moreover, our analysis focused on four specific models and a limited set of evaluation metrics. Future studies in this area should examine a wider range of models and include more metrics that capture diverse facets of QA performance, such as those designed for scientific or medical domains.

According to Chen et al. (2021), extensive, complex documents and inquiries requiring in-depth expertise are frequently challenging for current QA approaches to address. The complexity of real-world document sets, which often contain large amounts of text and require in-depth expertise, was not explored in our study. Addressing this limitation will be crucial for developing QA systems that can successfully handle such difficult scenarios.

5.2. Future Study

Future research should emphasize integrating qualitative evaluations alongside quantitative metrics, to provide a more comprehensive understanding of QA system performance and user experience. We recommend increasing the sample size and exploring a wider range of models and metrics to capture various aspects of performance.

Furthermore, integrating knowledge bases and textual inference can greatly improve the efficiency of the system and the accuracy of the evaluation. By highlighting these qualitative dimensions, researchers can obtain a greater understanding of user interactions and contextual challenges that QA systems face. This approach is consistent with recent developments in the field, such as the integration of large language models and the exploration of multi-modal QA systems.

This comprehensive approach not only contributes to academic discourse but also informs practical applications, ensuring that QA systems are tailored to meet real-world needs and complexities.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. *arXiv*, 1904.03323. https://doi.org/10.48550/arXiv.1904.03323

Alzubi, J. A., Jain, R., Singh, A., Parwekar, P., & Gupta, M. (2023). COBERT: COVID-19 question answering system using BERT. *Arabian Journal for Science and Engineering*, 48(8), 11003-11013. https://doi.org/10.1007/s13369-021-05810-5

Amur, Z. H., Kwang Hooi, Y., Bhanbhro, H., Dahri, K., & Soomro, G. M. (2023). Short-text semantic similarity (STSS): Techniques, challenges and future perspectives. *Applied Sciences*, 13(6), 3911. https://doi.org/10.3390/app13063911

Bahak, H., Taheri, F., Zojaji, Z., & Kazemi, A. (2023). Evaluating ChatGPT as a question answering system: A comprehensive analysis and comparison with existing models. *arXiv*, 2312.07592. https://doi.org/10.48550/arXiv.2312.07592

Baktash, J. A., & Dawodi, M. (2023). GPT-4: A review on advancements and opportunities in natural language processing. *arXiv*, 2305.03195. https://doi.org/10.48550/arXiv.2305.03195

Bowman, S. R., & Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding? *arXiv*,

2104.02145. https://doi.org/10.48550/arXiv.2104.02145

Campos-Mercade, P., Meier, A. N., Schneider, F. H., & Wengström, E. (2021). Prosociality predicts health behaviors during the COVID-19 pandemic. *Journal of Public Economics*, 195, 104367. https://doi.org/10.1016/j.jpubeco.2021.104367

Chen, X., Cui, Z., Zhang, J., Wei, C., Cui, J., Wang, B., Zhao, D., & Yan, R. (2021). Reasoning in dialog: Improving response generation by context reading comprehension. *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 35(14), 12683-12691. https://doi.org/10.1609/aaai.v35i14.17502

Chen, Y., & Eger, S. (2023). MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11, 804-825. https://doi.org/10.1162/tacl_a_00576

Deutsch, D., Bedrax-Weiss, T., & Roth, D. (2021). Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9, 774-789. https://doi.org/10.1162/tacl_a_00397

d'Hoffschmidt, M., Belblidia, W., Brendlé, T., Heinrich, Q., & Vidal, M. (2020). FQuAD: French question answering dataset. *arXiv*, 2002.06071. https://doi.org/10.48550/arXiv.2002.06071

Goodwin, T., Demner, D., Lo, K., Wang, L. L., Hersh, W. R., Dang, H. T., & Soboroff, I. M. (2020). *Epidemic question answering*. https://bionlp.nlm.nih.gov/epic_qa/

Herbold, S. (2024). Semantic similarity prediction is better than other semantic similarity measures. *arXiv*, 2309.12697. https://doi.org/10.48550/arXiv.2309.12697

Möller, T., Reina, A., Jayakumar, R., & Pietsch, M. (2020, July). COVID-QA: A question answering dataset for COVID-19. In K. Verspoor, K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, & B. Wallace (Eds.), *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics.

North, K., Zampieri, M., & Shardlow, M. (2023). Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9), Article 179. https://doi.org/10.1145/3557885

Pal, A. (2023). CLIFT: Analysing natural distribution shift on question answering models in clinical domain. *arXiv*, 2310.13146. https://doi.org/10.48550/arXiv.2310.13146

Raza, S., Schwartz, B., & Rosella, L. C. (2022). CoQUAD: A COVID-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC Bioinformatics*, 23(1), 210. https://doi.org/10.1186/s12859-022-04751-6

Soni, S., & Roberts, K. (2022). Toward a neural semantic parsing system for EHR question answering. *AMIA Annual Symposium Proceedings*, 2022, 1002-1011.

Turchin, A., Masharsky, S., & Zitnik, M. (2023). Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, 36, 101139. https://doi.org/10.1016/j.imu.2022.101139

Unanue, I. J., Parnell, J., & Piccardi, M. (2021). BERTTune: Fine-tuning neural machine translation with BERTScore. *arXiv*, 2106.02208. https://doi.org/10.48550/arXiv.2106.02208

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, December 4-9). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 1-11). Curran Associates Inc.

Wang, X., Liu, B., Xu, F., Long, B., Tang, S., & Wu, L. (2022). Feeding what you need by understanding what you learned. *arXiv*, 2203.02753. https://doi.org/10.48550/arXiv.2203.02753

Yang, A., Liu, K., Liu, J., Lyu, Y., & Li, S. (2018, July). Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. In E. Choi, M. Seo, D. Chen, R. Jia, & J. Berant (Eds.), *Proceedings of the Workshop on Machine Reading for Question Answering* (pp. 98-104). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2611

Zaib, M., Tran, D. H., Sagar, S., Mahmood, A., Zhang, W. E., & Sheng, Q. Z. (2021). BERT-CoQAC: BERT-based conversational question answering in context. In L. Ning, V. Chau, & F. Lau (Eds.), *Proceedings of the Parallel Architectures, Algorithms and Programming. PAAP 2020* (pp. 47-57). Communications in Computer and Information Science. https://doi.org/10.1007/978-981-16-0010-4_5

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *arXiv*, 1904.09675. https://doi.org/10.48550/arXiv.1904.09675

Call for Paper

Journal of Information Science Theory and Practice (JISTaP)

We would like to invite you to submit or recommend papers to Journal of Information Science Theory and Practice (JISTaP, eISSN: 2287-4577, pISSN: 2287-9099), a fast track peer-reviewed and no-fee open access academic journal published by Korea Institute of Science and Technology Information (KISTI), which is a government-funded research institute providing STI services to support high-tech R&D for researchers in Korea. JISTaP marks a transition from Journal of Information Management to an English-language international journal in the area of library and information science.

JISTaP aims at publishing original studies, review papers and brief communications on information science theory and practice. The journal provides an international forum for practical as well as theoretical research in the interdisciplinary areas of information science, such as information processing and management, knowledge organization, scholarly communication and bibliometrics.

We welcome materials that reflect a wide range of perspectives and approaches on diverse areas of information science theory, application and practice. Topics covered by the journal include: information processing and management; information policy; library management; knowledge organization; metadata and classification; information seeking; information retrieval; information systems; scientific and technical information service; human-computer interaction; social media design; analytics; scholarly communication and bibliometrics. Above all, we encourage submissions of catalytic nature that explore the question of how theory can be applied to solve real world problems in the broad discipline of information science.

Co-Editors in Chief : Gary Marchionini & Dong-Geun Oh

Please click the "Online Submission" link in the JISTaP website (http://www.jistap.org), which will take you to a login/account creation page. Please consult the "Author's Guide" page to prepare your manuscript according to the JISTaP manuscript guidelines.

Any question? Hyekyong Hwang (managing editor) : jistap@kisti.re.kr

JISTaP
Journal of Information Science Theory and Practice

Information for Authors

The Journal of Information Science Theory and Practice (JISTaP), which is published quarterly by the Korea Institute of Science and Technology (KISTI), welcomes materials that reflect a wide range of perspectives and approaches on diverse areas of information science theory, application and practice. JISTaP is an open access journal run under the Open Access Policy. See the section on Open Access for detailed information on the Open Access Policy.

A. Originality and Copyright

All submissions must be original, unpublished, and not under consideration for publication elsewhere. Once an article is accepted for publication, all papers are accessible to all users at no cost. If used for other researches, its source should be indicated in an appropriate manner and the content can only be used for uncommercial purpose under Creative Commons license.

B. Peer Review

All submitted manuscripts undergo a single-blind peer review process in which the identities of the reviewers are withheld from the authors.

C. Manuscript Submission

Authors should submit their manuscripts online via AccessON Peer Review Management System Plus (ACOMS^{+}). Online submission facilitates processing and reviewing of submitted articles, thereby substantially shortening the paper lifecycle from submission to publication. After checking the manuscript's compliance to the Manuscript Guidelines, please follow the "Online Submission" hyperlink in the top navigation menu to begin the online manuscript submission process.

D. Open Access and Creative Commons License Statement



All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. Open Access articles are automatically archived in the Korea Institute of Science and Technology Information (KISTI)'s Open Access repository (AccessON). All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/). Under this license, the authors retain full ownership of their work, while permitting anyone to use, distribute, and reproduce the content in any medium, as long as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

E. Manuscript Guidelines

Manuscripts that do not adhere to the guidelines outlined below will be returned for correction. Please read the guidelines carefully and make sure the manuscript follows the guidelines as specified. We strongly recommend that authors download and use the manuscript template in preparing their submissions.

1. Page Layout :

All articles should be submitted in single column text on standard Letter Size paper (21.59 × 27.94 cm) with normal margins.

2. Length :

Manuscripts should normally be between 4,500 and 9,000 words (10 to 20 pages).

3. File Type :

Articles should be submitted in Microsoft Word format. To facilitate the manuscript preparation process and speed up the publication process, please use the manuscript template.

4. Text Style :

- Use a standard font (e.g., Times New Roman) no smaller than size 10.
- Use single line spacing for paragraphs.
- Use footnotes to provide additional information peripheral to the text. Footnotes to tables should be marked by superscript lowercase letters or asterisks.

5. Title Page :

The title page should start with a concise but descriptive title and the full names of authors along with their affiliations and contact information (i.e., postal and email addresses). An abstract of 150 to 250 words should appear below the title and authors, followed by keywords (4 to 6).

Author1
Affilliation, Postal Address. E-mail
Author2
Affilliation, Postal Address. E-mail

ABSTRACT
A brief summary (150-250 words) of the paper goes here.

Keywords: 4 to 6 Keywords, separated by commas.

6. Numbered Type :

1. INTRODUCTION

All articles should be submitted in single column text on standard letter size paper (21.59 × 27.94 cm) with normal margins[1. Text should be in 11-point standard font (e.g., Times New Roman) with single line spacing. [1 Normal margin dimensions are 3 cm from the top and 2.54 cm from the bottom and sides.

2. SECTIONS

The top-level section heading should be in 14-point bold all uppercase letters.

2.1. Subsection Heading 1

The first-level subsection heading should be in 12-point bold with the first letter of each word capitalized.

2.1.1. Subsection Heading 2

The second-level subsection heading should be in 11-point italic with the first letter of each word capitalized.

7. Figures and Tables :

All figures and tables should be placed at the end of the manuscript after the reference list. To note the placement of figures and tables in text, "Insert Table (or Figure) # here" should be inserted in appropriate places. Please use high resolution graphics whenever possible and make sure figures and tables can be easily resized and moved.

Figure

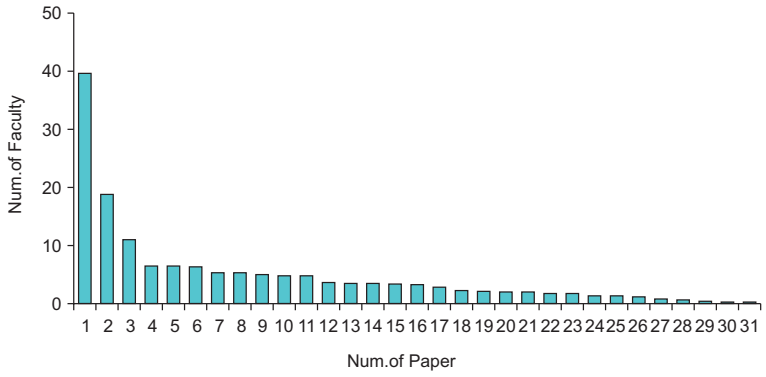


Fig. 1. Distribution of authors over publication count.

Table

Table 1. The title of table goes here

Study	Time period study	Data
Smith Wesson (1996)	1970 - 1995	684 papers in 4 SSCI journals
Reeves [a (2002)	1997 - 2001	597 papers in 3 SSCI journals
Jones Wilson [b (2011)	2000 - 2009	2,166 papers in 4 SSCI journals

[a Table footnote a goes here
[b Table footnote b goes here

8. Acknowledgements :

Acknowledgements should appear in a separate section before the reference list.

9. Conflicts of Interest :

Authors of manuscripts must disclose any potential conflicts of interest at the time of submission. Statements on conflict of interest have no influence on the editorial decision to publish.

10. Citations :

Citations in text should follow the author-date method (authors' surname followed by publication year).

- Several studies found... (Barakat et al., 1995; Garfield, 1955; Meho & Yang, 2007).
- In a recent study (Smith & Jones, 2011)...
- Smith and Jones (2011) investigated...

11. Reference List :

Reference list, formatted in accordance with the American Psychological Association (APA) style, should be alphabetized by the first authors last name.

Journal article

- Author, A., Author, B. & Author, C. (Year). Article title. *Journal Title*, volume(issue), start page-end page.
- Smith, K., Jones, L. J., & Brown, M. (2012). Effect of Asian citation databases on the impact factor. *Journal of Information Science Practice and Theory*, 1(2), 21-34.

Book

- Author, A., & Author, B. (Year). *Book title*. Publisher Name.
- Smith, K., Jones, L. J., & Brown, M. (2012). *Citation patterns of Asian scholars*. Sage.

Book chapter

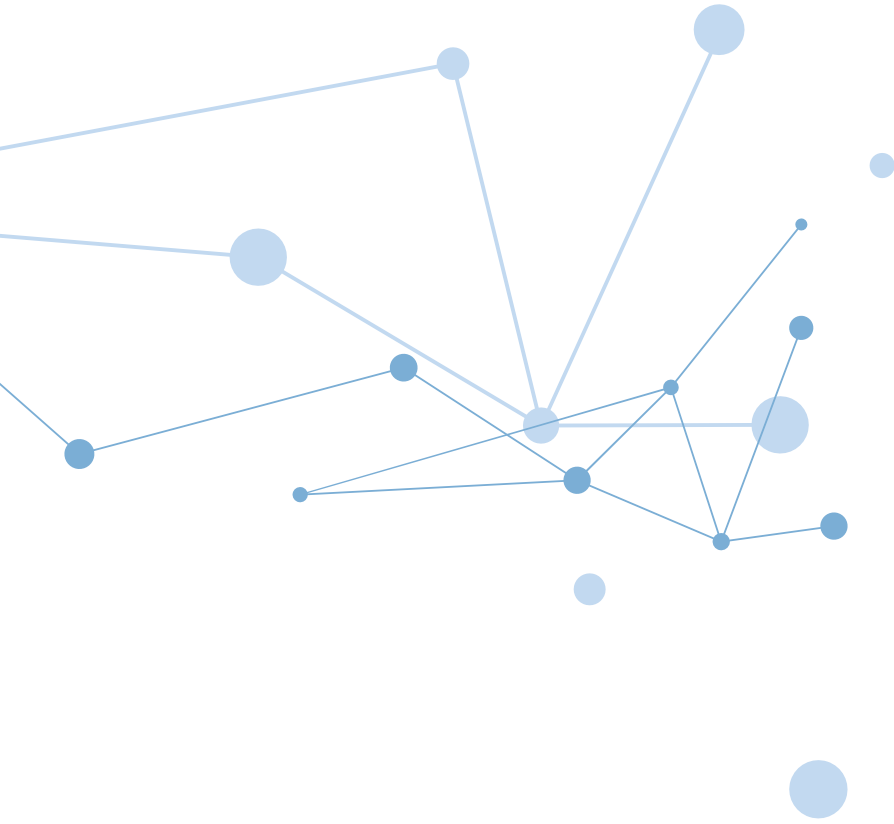
- Author, A., & Author, B. (Year). Chapter title. In A. Editor, B. Editor, & C. Editor (Eds.), *Book title* (pp. xx-xx). Publisher Name.
- Smith, K. & Brown, M. (2012). Author impact factor by weighted citation counts. In G. Martin (Ed.), *Bibliometric approach to quality assessment* (pp. 101-121). Springer.

Conference paper

- Author, A., & Author, B. (Year, Month Day). Article title. In A. Editor & B. Editor (Eds.), *Conference title* (pp. xx-xx). Publisher Name.
- Lin, H., Sun, B., Wu, J. & Xiong, H. (2016, June 24-26). Topic detection from short text: A term-based consensus clustering method. In B. Yang (Ed.), *2016 13th International Conference on Service Systems and Service Management (ICSSSM 2016)* (pp. 1-6). IEEE.

Online document

- Author, A., & Author, B. (Year). *Article title*. URL.
- Smith, K. & Brown, M. (2010). The future of digital library in Asia. *Digital Libraries*, 7,111-119. http://www.diglib.org/publist.htm.



JISTaP

Journal of Information
Science Theory and Practice

KISTI www.kisti.re.kr
66, Hoegi-ro, Dongdaemun-gu, Seoul, Republic of Korea
(T) +82-2-3299-6105 (F) +82-2-3299-6067
https://www.jistap.org