eISSN : 2287-4577
pISSN : 2287-9099
http://www.jistap.org
Vol. 9 No. 1 March 30, 2021

Journal of Information Science Theory and Practice

Journal of Information Science Theory and Practice

Vol. 9 | No. 1 | March 30, 2021

JISTaP

01 Using Practice Context Models to Knowledge Management in Proof-of-Concept Activities: A Contribution of Knowledge Networks and Percolation Theory
24 Identification of Profane Words in Cyberbullying Incidents within Social Networks
35 Topic Modeling and Sentiment Analysis of Twitter Discussions on COVID-19 from Spatial and Temporal Perspectives
54 ISRI - Information Systems Research Constructs and Indicators: A Web Tool for Information Systems Researchers
68 The Use of Weblogs as a Tool for Thai Political Engagement
79 Content Analysis of the Facebook Pages of Selected Academic Libraries in Vietnam

KiSTi
www.kisti.re.kr

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

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 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

Publisher

Korea Institute of Science and Technology Information
66, Hoegi-ro, Dongdaemun-gu, Seoul, Republic of Korea
(T) +82-2-3299-6102
(F) +82-2-3299-6067
E-mail: jistap@kisti.re.kr
URL: http://www.jistap.org

Managing Editor: Ji-Young Kim, Eungi Kim

Copy Editor: Ken Eckert

Design & Printing Company: MEDrang Inc.

31st Floor, Prime Center, 85 Gwangnaru-ro 56-gil, Gwangjin-gu, Seoul 05116, 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 au-thors, permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far 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).

2021 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

Associate Editors

Kiduk Yang
Kyungpook National University, Korea

Taesul Seo
Korea Institute of Science and Technology Information, Korea

Managing Editors

Ji-Young Kim
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

Lemen Chao
Renmin University of China, China

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

Helen Partidge
University of Southern Queensland, Australia

M. P. Satija
Guru Nanak Dev University, India

Ou Shiyan
Nanjing University, China

Paul Solomon
University of South Carolina, USA

Wolfgang G. Stock
Heinrich-Heine-Universität Düsseldorf, Germany

Consulting Editors

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

Tsutomu Shihota
St. Andrews University, Japan

Guancan Yang
Renmin University of China, China

2021 Copyright © Korea Institute of Science and Technology Information

JISTaP
Journal of Information Science Theory and Practice

Table of Contents

Vol. 9 No. 1 March 30, 2021
Journal of Information Science Theory and Practice • http://www.jistap.org

Articles

01 Using Practice Context Models to Knowledge Management in Proof-of-Concept Activities: A Contribution of Knowledge Networks and Percolation Theory
Antonio Jose Rodrigues Neto, Maria Manuel Borges, Licinio Roque

24 Identification of Profane Words in Cyberbullying Incidents within Social Networks
Wan Noor Hamiza Wan Ali, Masnizah Mohd, Fariza Fauzi

35 Topic Modeling and Sentiment Analysis of Twitter Discussions on COVID-19 from Spatial and Temporal Perspectives
Iyad AlAgha

54 ISRI - Information Systems Research Constructs and Indicators: A Web Tool for Information Systems Researchers
João Varajão, António Trigo, Tiago Silva

68 The Use of Weblogs as a Tool for Thai Political Engagement
Sutthinan Chuenchom

79 Content Analysis of the Facebook Pages of Selected Academic Libraries in Vietnam
Duong Thi Phuong Chi

Call for Paper

Information for Authors

2021 Copyright © Korea Institute of Science and Technology Information

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

RESEARCH PAPER
J Inf Sci Theory Pract 9(1): 1-23, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.1.1

Using Practice Context Models to Knowledge Management in Proof-of-Concept Activities: A Contribution of Knowledge Networks and Percolation Theory

Antonio Jose Rodrigues Neto*
University of Coimbra, Centre of 20th Century Interdisciplinary Studies (CEIS20), Faculty of Arts and Humanities, Coimbra, Portugal
E-mail: neto@student.uc.pt

Maria Manuel Borges
University of Coimbra, Centre of 20th Century Interdisciplinary Studies (CEIS20), Faculty of Arts and Humanities, Coimbra, Portugal
E-mail: mmb@fl.uc.pt

Licinio Roque
University of Coimbra, Centre for Informatics and Systems of the University of Coimbra (CISUC), Department of Informatics Engineering, Coimbra, Portugal
E-mail: lir@dei.uc.pt

ABSTRACT

This study introduces novel research using Practice Context Models supported by Knowledge Networks and Percolation Theory with the aim to contribute to knowledge management in Proof-of-Concept (PoC) activities. The authors envision this proposal as a potential instrument to identify network structures based on a percolation (propagation) threshold and to analyze the importance of nodes (e.g., practitioners, practices, competencies, movements, and scenarios) during the percolation of knowledge in PoC activities. After thirty months immersed in the natural PoC habitat, acting as observers and practitioners, and supported by an ethnographic exercise and a designer-research mindset, the authors identified the production of meaning in PoC activities occurring in a hermeneutic circle characterized by the presence of several knowledge networks; thus, discovering the 'natural knowledge' in PoC as a spectrum of cognitive development spread throughout its network, as each node could produce and disseminate certain knowledge that flows and influences other nodes. Therefore, this research presents the use of Practice Context Models 'connected' to Knowledge Networks and Percolation Theory as a potential and feasible proposal to be built using the attribution of values (weights) to the nodes (e.g., practitioners, practices, competencies, movements, scenarios, and also knowledge) in the context of PoC with the aim to allow the players (e.g., PoC practitioners) to have more flexibility in building alliances with other players (new nodes); that is, focusing on those nodes with higher value (focus on quality) in collaboration networks, i.e., alliances (connections) with the aim to contribute to knowledge management in the context of PoC.

Keywords: proof-of-concept practices, practices, context model, knowledge networks, percolation theory, knowledge management

Received: December 5, 2020 **Accepted:** February 2, 2021

***Corresponding Author:** Antonio Jose Rodrigues Neto
https://orcid.org/0000-0003-2364-0927
E-mail: neto@student.uc.pt

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far 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.

© Antonio Jose Rodrigues Neto, Maria Manuel Borges, Licinio Roque, 2021

1. INTRODUCTION

Knowledge is one of the basic factors for a country's scientific and technological development (Kuramoto, 2006), which is affirmed by Borges (2006) who notes that science shapes the way we perceive the world when determining various technical, economic, and social instances.

According to Jacobetty (2010), science can be described as being composed of a set of systematic research and investigation practices, with the objective of producing a specific type of knowledge. Thus, in the search for knowledge about reality, the achievement of scientific objectivity is operated in a community-based way through increasingly vast networks of densely-related observers in a universe of collaborations stimulated by improvements in means of transport and communication. However, there are many obstacles, such as the lack of information and the existence of barriers to accessing data, among others, but there is still space to find motivation in the development of science and practices and, in our opinion, one such motivation is the **ceaseless search for the generation of knowledge** (our emphasis), especially within Proof-of-Concept (PoC) activities (Neto et al., 2018, 2019).

1.1. Demystifying PoC Activity

The applicability of PoC activities is a common practice in several areas of knowledge and can be used by different organizations for several purposes, including the evaluation of new technology products or even in the pharmaceutical industry, where the evaluation of a new drug (e.g., a vaccine) requires a period of validation of its effectiveness, which is nothing more than a PoC, or Proof of Principle, according to Schmidt (2006).

In the context of our research, we acknowledge PoC as an activity with a set of practices performed by its practitioners and consumed by organizations of validation and experimentation of products and technologies with the aim of contributing to the construction and dissemination of knowledge in the study and understanding of the performance of certain artifacts and their phenomena in a given area of knowledge (Neto et al., 2020b). Therefore, the PoC activity stimulates the production and consumption of knowledge of practitioners and organizations, as well as their communities of practice, which will typically **generate tacit knowledge** (our emphasis) (Polanyi, 2009) of their applications and business processes, where in this context we understand knowledge as "a fluid mix of framed experience, values, contextual information, [and] this becomes embedded not only in documents or repositories but also in organizational routines, processes, **practices**, and norms" (our emphasis) (Davenport & Prusak, 1998, pp. 5-6).

On the other hand, we recognize that the lack of characterization (contextualization) of practices in the context of PoC (Kendig, 2016; Neto et al., 2018), as well as a model of the context of these practices, can compromise the formation of the interpretations and understandings of its practitioners with regard to the knowledge production and dissemination as we envisage a PoC activity composed of a **set of tacit and explicit movements** (our emphasis) by its practitioners, where "these movements may constitute several practices practitioners perform across organizational networks, for learning about, experimenting and evaluating new products or technologies, in the domain of IT [Information Technology]" (Neto et al., 2020b, p. 1).

According to Kendig (2016, p. 735), PoC is a term

> frequently used in descriptions of research sought in program announcements, in experimental studies, and in the marketing of new technologies. It is often coupled with either a short definition or no definition at all, with **its meaning assumed to be fully understood. This is problematic** (our emphasis).

Neto et al. (2020b, pp. 1-2) describe several PoC studies in the domain of Information Technology (IT) and within their final results those authors demonstrate the need for further contextualization, as well as more grounded and detailed analysis of practices rather than success or failure of experiments, e.g., such as in a proposed block diagram model in sequence by Barnes et al. (2009) or simple methods. In other words, in the context of PoC, we do not envisage PoC practitioners' movements "unfolding always in a well-defined sequence" (Neto et al., 2020b, p. 12).

In the context of PoC, we do not see these movements taking place in a well-defined mode of operation that would allow certain procedures and results to be carried out in a consistent manner. On the other hand, we do visualize a set of free and specialized movements that, in some way, comprise a PoC practice, being carried out by their practitioners in the development and execution of this activity, so that they can contribute to the exploration and reflection of the behavior and performance of the technological artifacts and the phenomena under study. Thus, we question the construction of PoC being seen as a "cake recipe."

That is, we understand a "cake recipe" to mean a set

 https://doi.org/10.1633/JISTaP.2021.9.1.1

of predefined instructions that, if strictly followed by its practitioner (that is, whoever makes the cake), its (final) result will always be the same (that is, the cake), regardless of the practitioner. We understand "cake recipe" to be a definition that does not fit well with the demanding context of PoCs, which more often than not resemble a convoluted research process. In addition, in the context of PoCs, we highlight the discovery, exploration, and reflection of several unexpected, unknown, unexplored, and even enigmatic phenomena during the development and execution of this activity, and we question the effectiveness of using simple recipes in the context of PoC.

Thus, we envisage that some of the PoC activities could potentially be framed (developed and executed) using agile and simple methods; however, based on our experience as practitioners and observers in the natural PoC habitat, we already noted a trend of PoC activities in the IT world, especially related to digital transformation initiatives by large organizations, which become increasingly complex and lengthy, especially due to the rapidity and complexity of technological evolution, such as with Artificial Intelligent (AI) PoCs, Cloud Computing PoCs, Containerization PoCs, and Data Storage Performance Characterization PoCs, among others.

In the context of PoC, we perceive a relevant phenomenon that exists, and is implicit in this context, involving its practitioners, the organizations, and the artifacts under study. This phenomenon represents one or more ways of seeing, interpreting, and doing things in that world, resulting in development problems in a sociotechnical context, in addition to increasing the likelihood of compromising the reliability, reproducibility, and reuse of knowledge in this activity, which may affect the proper use of that knowledge by organizations, PoC practitioners, and their community of practices.

Thus, when these practitioners respond to different areas of practice, whether indeterminate or unknown, they have several conversations (dialogues) of a reflective characteristic with the materials of their situations, such as, for example, exploration and reflection on other complex worlds (organization world), such as other information systems and their artifacts. In other words, PoC practitioners tend to remake part of their practical world (PoC), which may still reveal a **tacit set of new practices** (our emphasis) and interactions in the context of PoC. That is, a new construction and vision of this practical world, in other words, the development and execution of a PoC, on which all its practices will be based, contributes to the construction and dissemination of knowledge for the activity and, in some way, contributes to the construction of a knowledge base that can influence new constructions of new activities, which by analogy, we envisage as the construction of a large and complex knowledge network.

Hence, we can expect an improvement of the whole practice through accumulated practice knowledge and vice versa during those PoC activities. Thus, the conclusion is that we should aim to learn to recognize dialectic interactions between the practices in co-evolution and the activity itself, thus creating knowledge and transforming the activity system as a whole. This continuous and circular aspect of learning (a network) through emergent PoC practice emphasizes the critical role of sociotechnical movements such as dialogues, reflections, improvisations, and negotiations.

From a practitioner perspective, **PoC activities are anything but static** (our emphasis); rather, they are experiences that are forever evolving, where each activity has its own history, embedding 'past' phases and 'future' client expectations. Hence, we envision activity structure in a PoC context as that which is conceived of at a given moment, and it then undergoes a process of evolution, where old and new practices cohabitate, and they might be recreated and transformed so that the activity reaches its pragmatic 'format.' Activities also evolve with the improvement of mediators and object definition, inducing changes in PoC practices. Thus, we understand an important factor of productivity and competitiveness for organizations: the ability of individuals, and the organizations themselves, to create, process, and transform information and knowledge during the execution of PoC.

1.2. Proposed PoC Practices Context Model

With the aim to contribute to PoC activities, their practitioners, and their community of practice, we identified ten (10) practices in the context of PoC, within which those practices are characterized as a cycle of knowledge production (Neto et al., 2019). In other words, we identified and characterized how these practices contribute to PoC development, by mapping them in Information Systems Development problems within the Context Engineering framework proposed by Roque (2004).

Further, we developed a PoC practices context model (Fig. 1) (Neto et al., 2020b) to represent practices as an emergent sociotechnical reality, a representation of the PoC context and its practices as ongoing forms of activity, with the aim of making PoC observable and an object of explicit reasoning among practitioners and their communities of practice.

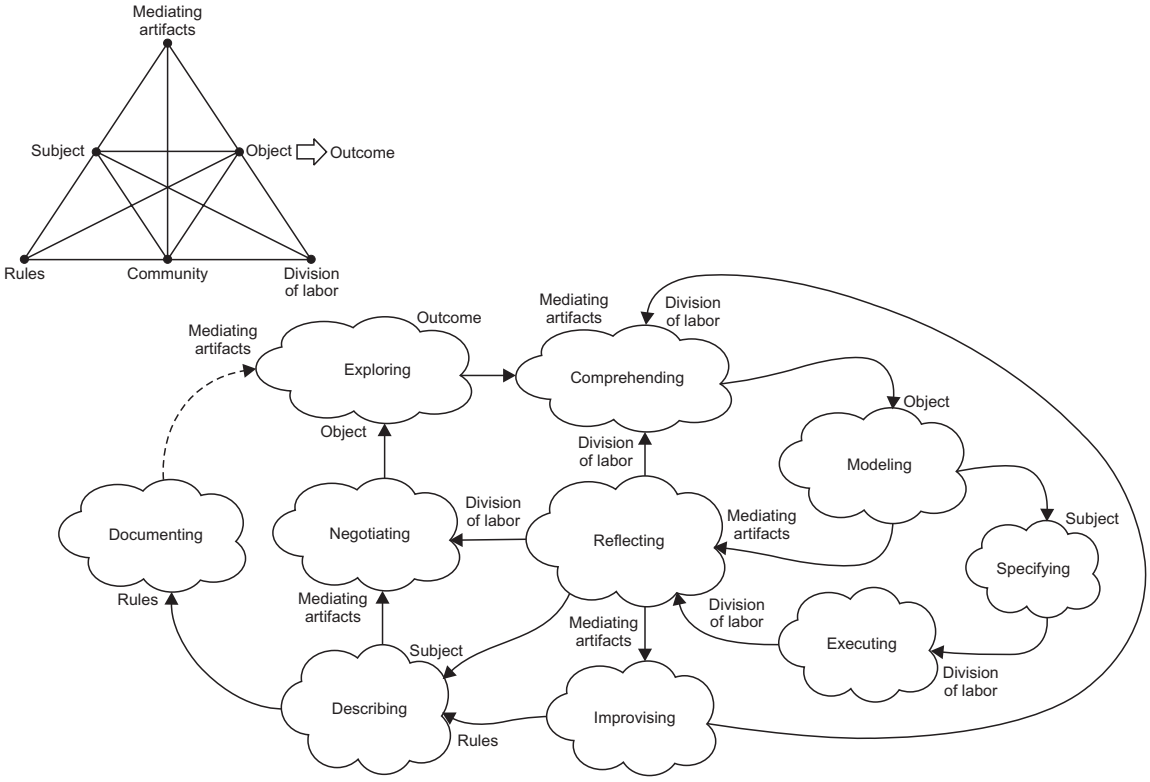


Fig. 1. Proof-of-Concept practices context model (Neto et al., 2020b).

We constructed the PoC practices context model in aiming to contribute to the discovered interactions between practices, thus fostering their co-evolution, and we frame these practices in the context of activity systems, grounded in sociotechnical phenomena, and used Activity Theory (Engeström, 1987) to substantiate our reflections on those practices and map mediators flowing between them. In other words, we position this as a language contribution to improve reflection on practice and to further enable the study of the role of emergent practices in system design.

In other words, the set of practices represented in the context model organize a set of relational flows or paths enabling the realization of a diversity of practice patterns, contingent with each PoC activity demands, in which the

> PoC knowledge is, thus grounded in a relationship among its practices, engineering the sociotechnical context that invokes its use. By collecting and assembling fragments of practitioners' movements in context, we highlight the importance of developing this PoC context model to better understand their role in the overall activity system (Neto et al., 2020b, p. 12).

Thus, we aimed to propose a PoC practices context model with respect to how those practices interact and evolve in the PoC context or how they present a relationship between a learning process in a PoC context and the Theory of Expansive Learning (Engeström, 1987, 2001, 2016). In other words, PoC context models are viewed as activity systems that are

> increasingly interconnected and interdependent, and many recent studies of expansive learning take as their unit of analysis a constellation of two or more activity systems that have a partially shared object. Such interconnected activity systems may form a producer-client relationship, a partnership, **a network** or some other pattern of multi-activity collaboration (our emphasis) (Engeström, 2016, p. 45).

Note that the above research (Neto et al., 2020b) presents a Practice Context Model in the context of the development and execution of PoC activities, whereby the goal with this proposed model was to provide a context model with the aim of acting as a compass and to 'map' those movements, whether they are acting currently or will be acting in the future, so that its practitioners can reflect on

where they are and what they have done in order to take their development goals further (Roque et al., 2004).

1.3. Knowledge Networks in the Context of PoC

During the development of the PoC practices context model (Neto et al., 2020b), we learned that knowledge is not only produced and disseminated at the end of the PoC activity, but **during all the interactions** (our emphasis) between all the elements in a PoC activity (Neto et al., 2019), i.e., PoC scenarios, movements, competencies, and practices in the context of PoC), which we describe as:

- **Movements:** the actions-codes performed by PoC practitioners, regarding knowledge construction and dissemination during the PoC activity, e.g., some dialogues in a sociotechnical context
- **Competencies:** a condensed set of skills of the PoC practitioner, developed from the range of identified movements (action-codes)
- **Scenarios:** a set of dialogues and sketches that happened during the PoC activities
- **Practices:** In the context of our research, we have adopted the term of *practice* (Neto et al., 2018, p. 271) in accordance with the works of Isabelle et al. (2012) and Schön (1983). Isabelle et al. define practices as "both what people do in situations and the way they do it [in those situations]. Practices are knowing in action (Piaget, 1974), that is, [practices are] dispositions enacted in specific situations (Bourdieu, 1990)" (2012, p. 476). Furthermore, Schön notes, "I offer an approach to epistemology of practice based on a close examination of what some practitioners—architects, psychotherapists, engineers, planners, and managers—actually do" (Schön, 1983, p. viii).

We also learned that the practitioner is also not an isolated individual in a PoC context, for "the practitioner lives in a given context and interacts with that context; at the same time, the PoC practitioner influences and is influenced by the context" (Neto et al., 2020a, p. 130), a definition that strengths the hermeneutic nature in the PoC context (Neto et al., 2020a). In other words, we envisage the PoC context to be represented through a network composed of several elements (nodes), that is, the practitioners, their movements, their competencies, the PoC scenarios, and even information and knowledge produced and disseminated during the development and execution of the PoC, and their connections. For example, node A (that is, a practitioner in the context of PoC) connects to node B (that is, a practice), where that connection (link), in some way, produces knowledge and potentially disseminates it to other nodes in the PoC Knowledge Network (see Fig. 2).

Thus, according to Vygotsky, knowledge is "constructed as a result of the personal and subjective experience of an activity" which precedes knowledge, and "the technologies themselves are artifacts of practical activity… [Thus, when] the artifacts change, so does the activity [involving] new learning cycles" (Fino, 2001, p. 274), in a *continuum* learning cycle, i.e., the production and dissemination of knowledge in a PoC context is grounded by a complex knowledge network, which in this study we term as **PoC Knowledge Networks** (our emphasis).

As can be seen in Fig. 2, we present a knowledge network based on a PoC practices context model, where practitioners are connected in the practices, as well as the competencies, movements, and PoC scenarios (please refer to the Appendix for more examples). Thus, each practice produces and disseminates knowledge, which we envisage as another node in the knowledge network.

Thus, in the context of our research, we recognize a network in the context of PoC in the same sense as Menczer et al. (2020, p. 323), that is, networks are one way "to represent and study simple and complex interactions." The same authors present a network, that is,

> the simplest description of a set of interconnected entities, which we call nodes, and their connections, which we call links. The network representation is so general and powerful because it strips out many details of a particular system and focuses on the interactions among its elements (Menczer et al., 2020, p. 325).

Also, Barabási (2016, p. 45) describes a network as "a catalog of a system's components often called nodes or vertices and the direct interactions between them, called links or edges." That is, nodes represent the elements (components) in the system, e.g., the practices in the context of PoC, and the interactions between these practices are called links or edges. In other words, each practitioner who is 'connected' within a practice produces as an outcome some type of knowledge (conscious or unconscious) that serves as an additional connection to the next element in the network, thus forming (developing) a knowledge network (or several networks) in the context of PoC.

Thus, one approach is to reflect on the PoC practices context model as a complex network (graph) for knowl-

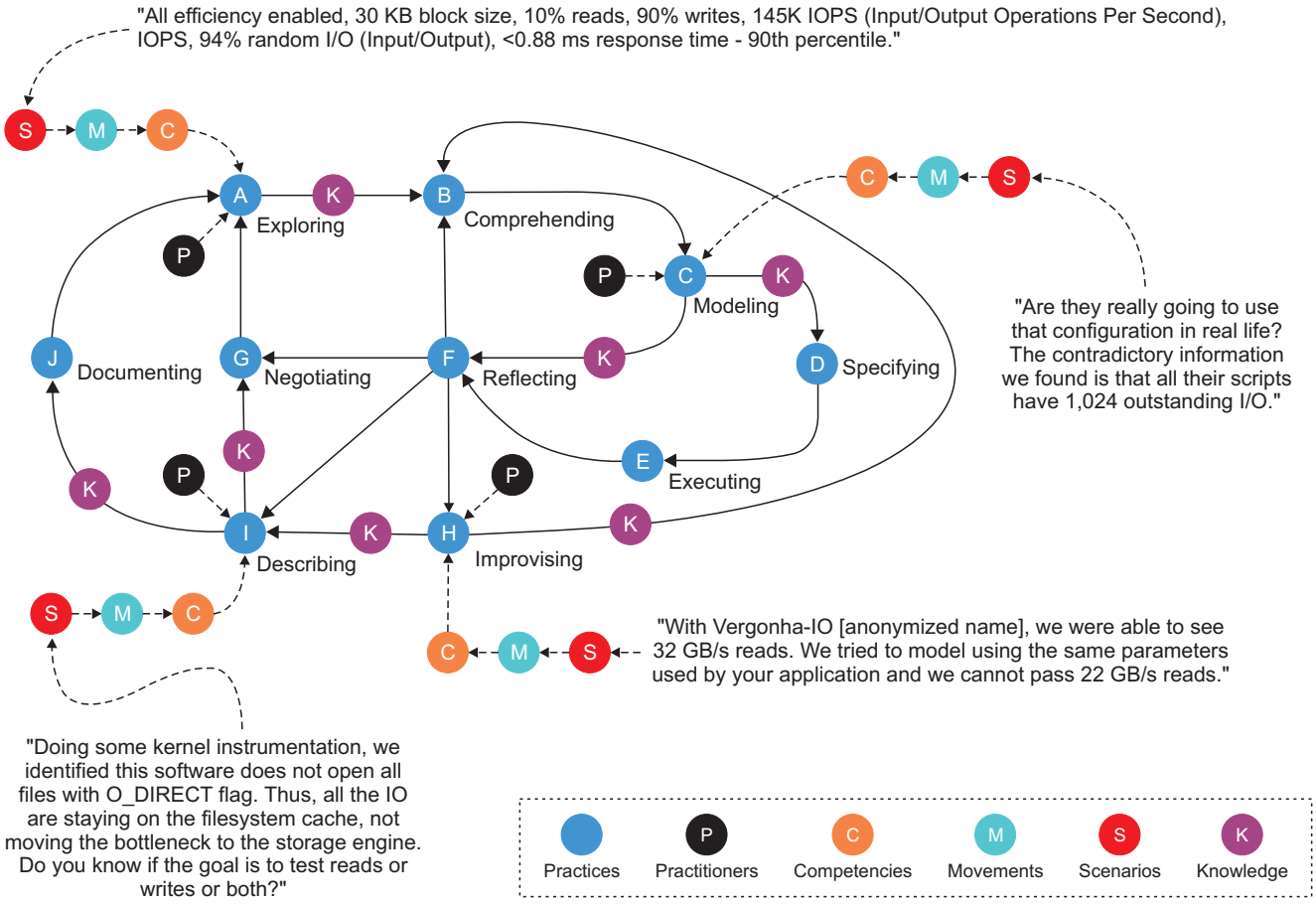


Fig. 2. Proof-of-Concept Knowledge Network based on Practice Context Model (Neto et al., 2020b).

edge production and the impact of one node on another, with the aim to contribute to the understanding of practitioners, regarding where and how they can improve in relation to practices in the context of PoC, through the construction of these collaborative networks.

We observed that the PoC Knowledge Network (Fig. 3) is formed by several networks (subnets), where each subnet (or cluster) aims to generate knowledge while employing a reuse protocol and open characteristics, i.e., aiming to allow other networks to use each of the networks.

As can be seen in Fig. 3, players 1 and 2 (PoC practitioners) form a network (Network A) during the practice of Exploring. This network produces some knowledge (Knowledge A), which can form a new vertex (node) for any other network. We can observe the same operation in the practice of Documenting and Improvising, the formation of networks called Network U and Network Y, respectively. We also can observe the creation of a network (Network n) composed of two subnets (Network B with players 4 and 5 and Network C with players 1 and 3) through the practice of Describing.

However, the subnet Network C produces some specific knowledge (Knowledge G) which may or may not have influenced the construction of knowledge for the entire network. In other words, the product of a network is the knowledge produced, which may or may not be reused by other networks. The networks (graph) represented in Fig. 3 can exhibit a high degree of heterogeneity, with fluctuations in connectivity parameters that span several orders of magnitude (see later in Fig. 4).

1.4. A Borrowed Theory from Physics: Introducing Percolation Theory in the Context of PoC

To the best of our knowledge, over the past two decades, Percolation Theory (Sahimi, 1994; Stauffer & Aharony, 1992) has been used to explain and model a wide variety of phenomena that are of industrial and scientific importance. According to Sahimi (1994, abstract section), examples include "characterization of porous materials and reservoir rocks, fracture patterns and earthquakes

 https://doi.org/10.1633/JISTaP.2021.9.1.1

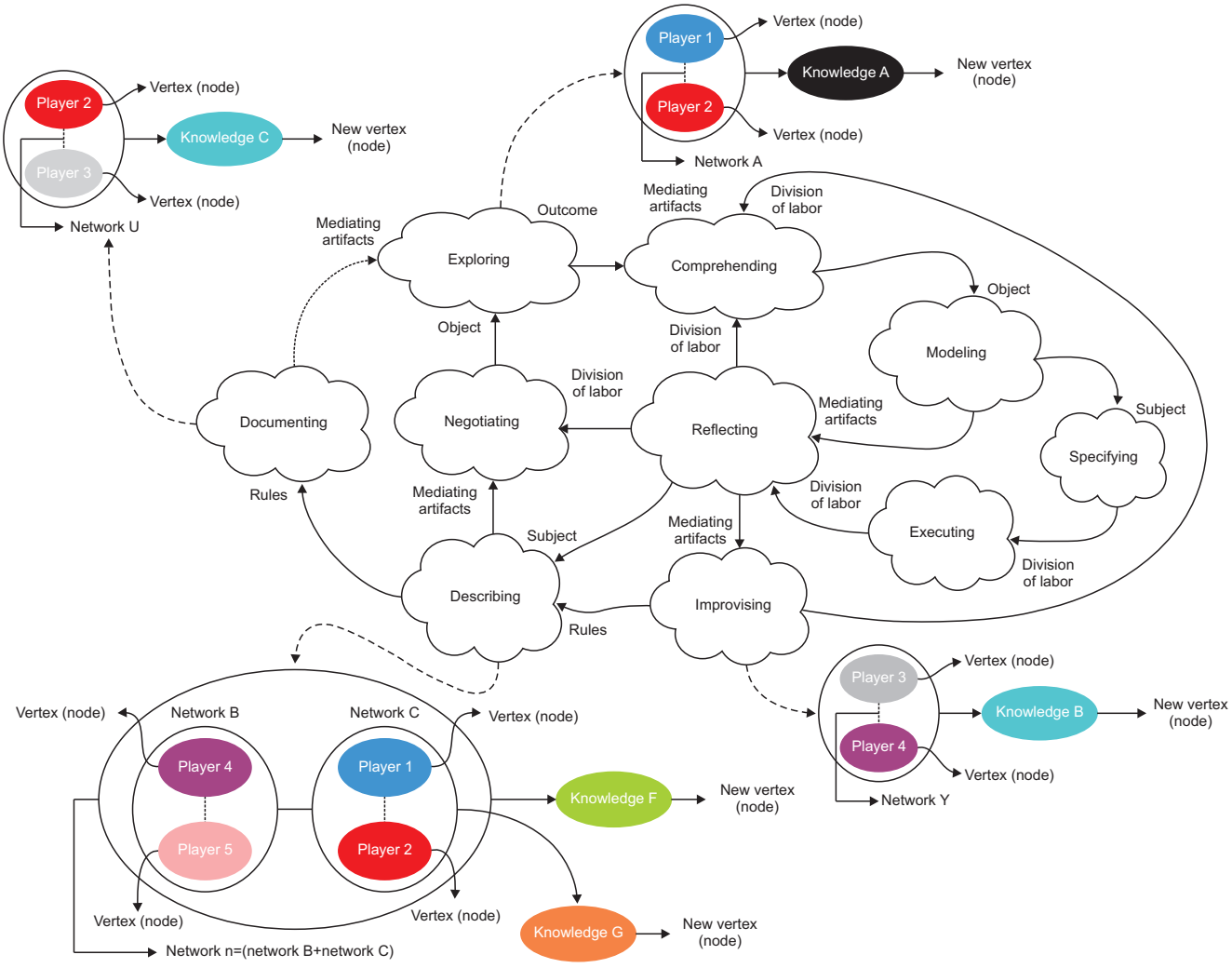


Fig. 3. Proof-of-Concept Knowledge Network and its sub-networks.

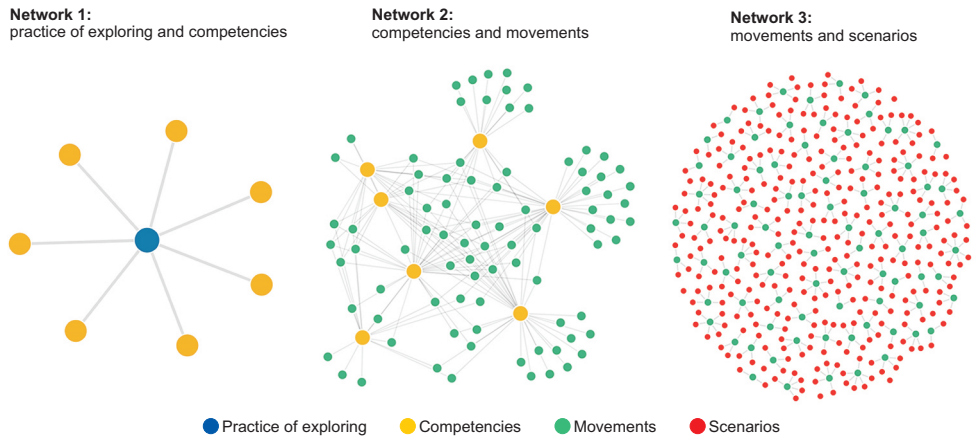


Fig. 4. Several Proof-of-Concept Knowledge Networks based on the practice of Exploring.

in rocks, calculation of effective transport properties of porous media permeability, conductivity, diffusivity, etc., groundwater flow, polymerization and gelation, biological evolution, galactic formation in the universe, spread of knowledge, and many others."

Also, according to Piraveenan et al. (2013, p. 1), Per-

colation Theory is "attractive because it provides connections to several well-known results from statistical physics, in terms of percolation thresholds, phase transitions, long-range connectivity, and critical phenomena in general." However, several studies, such as those of Popescul (2012) and Block et al. (2015) present a **new way of thinking about Percolation Theory** (our emphasis).

According to Popescul (2012, p. 445), Percolation Theory promotes "thinking regarding the relation between innovation and knowledge using a model borrowed from Physics, in aiming to prove whether knowledge resources can 'flow' (be percolated) in a network or a grid, in order to be transformed into technological innovation." Block et al. (2015, p. 23) note that "the main specificity of percolation is connectedness, as percolation theory is a general mathematical theory of connectivity. In physics, percolation processes result from the counteraction of two forces – connectivity and receptivity" and discuss their research as a "model of interaction within the company [that] is built on the percolation processes model, which was developed in Physics and represents a universal mathematical apparatus for researching clustered and non-clustered environments."

In the context of our research, we intend to reflect on the possibility of using Percolation Theory in conjunction with Knowledge Networks in aiming to contribute to the identification and analysis of the number of connections (links) between nodes in the knowledge network, and as a way of 'measuring' the capacity of these nodes regarding the dissemination (propagation) of information and knowledge through a PoC network. Thus, we expect to know how properties such as connectivity and receptivity of the knowledge network can be used to explain the process of knowledge sharing within an organization that relies on several PoC activities, with the aim to contribute to their digital transformation and technology innovation.

In other words, we aim to provide a theoretical and novel proposal based on Practice Context Models supported by Knowledge Networks and Percolation Theory, aiming to acknowledge the creation of complex networks with strategic and collaborative characteristics for the production and dissemination of knowledge, and contribute to knowledge management in PoC activities.

2. RESEARCH QUESTION

The motivation for our research was provided by the opportunity to propose a novel perspective on how to manage knowledge in the context of PoC, with the aim to contribute to knowledge management in PoC activities and provide a distinct reflection on how the knowledge can be constructed, disseminated, and potentially measured, through Practice Context Models with the support of Knowledge Networks and Percolation Theory.

According to Fanfan (2012, p. 118), identifying and distinguishing knowledge – knowledge management – is "any process or practice of creating, acquiring, capturing, sharing, using and evaluating knowledge wherever it resides." However, one question has motivated us to continue and advance our research in the context of PoC, thus, we start our research with the following question: "How can we 'manage knowledge' in the context of PoC?"

We understand that in order to present a proposal with the aim to contribute to knowledge management in PoC activities, initially it is necessary to define a networking core model to be used, identify and classify its nodes and connections, and later, construct a model with the ability to "measure" these nodes.

In other words, we need to define (1) a Practice Context Model to be the core for the knowledge network, (2) identify and classify the nodes and their connections in that knowledge network, and (3) present a conceptual proposal with the aim to contribute to the evaluation of these nodes. Therefore, we present our two main objectives in this research:

- **Objective 1:** identify, classify, and map the connections of the elements (PoC scenarios, movements, competencies, and practices) in a PoC context with the support of Knowledge Networks
- **Objective 2:** a proposal to "measure" the nodes (PoC elements) distributed in those knowledge networks with the support of Percolation Theory

As can be seen in Fig. 1, we aim to reflect on the PoC practices context model (Neto et al., 2020b) as our network core, where each practice is represented in a cloud (e.g., Reflecting) and the set lines represent the interactions (flows) between those practices, which would translate to some way of transporting information and knowledge and its ways of obtaining among those interactions. In other words, in the context of PoC, each activity structure relates practitioners to an object or motive of each activity in the PoC context.

We envisage PoC practices playing an important role in mediating between practitioners (as subjects) and how they organize, regulate, and use instruments to achieve the PoC result. Through these mediations, results must be

 https://doi.org/10.1633/JISTaP.2021.9.1.1

obtained and knowledge about the object of the PoC gets produced. Thus, turning the PoC object into knowledge is a shared motive across all PoC interactions (flows or connections in a network), and not only at the end of the PoC activity.

Due to the importance of knowledge, we understand that it is necessary to change our thinking with regards to organizational, strategic, and competitive innovation in organizations (Klaus et al., 2016), especially in how to execute and view PoC activities. From another perspective, the creation of the organizational knowledge is consolidated, in terms of organizational processes, with the purpose of assisting and amplifying the creation of knowledge, through the individuals in the organization, as part of their knowledge networks. Through this line of thought, it is possible to identify and distinguish between the different levels of social interaction in a PoC network, in which knowledge has been transformed and recognized, i.e., identifying and distinguishing the different levels of knowledge in the practices of the PoC context.

We also strengthen our understanding (vision) that PoC can be understood as "forming a cycle of dialectical reproduction of activities oriented towards the enaction of a body of knowledge, based on the interpretation and comprehension of all parts" (Neto et al., 2020a, p. 130) by its practitioners, which emphasizes the hermeneutic nature of the PoC context which, according to Gadamer (2013), is the philosophical theory of knowledge which states that "all cases of understanding necessarily involve both interpretation and application; a definition that we emphasize and which we consider to be lacking in the reflection and consciousness of PoC practitioners" (Neto et al., 2020a, p. 129).

3. RESEARCH METHOD

In this study, we aim to contribute to the body of scientific literature with regard to how the PoC practices context model (Neto et al., 2020b), with the support of Knowledge Networks and Percolation Theory, can contribute to knowledge management in PoC activities.

Thus, we based our research method on the characteristics of qualitative research acting directly within the regular environment of the PoC practitioners (which we term as a natural PoC habitat), where there is a direct source of data and the researcher becomes an instrument of research. Thus, the PoC environment and PoC practitioners were holistically observed and analyzed, while not being reduced to variables; rather, they were observed as a whole.

In other words, we describe our research method as an interpretive (qualitative) paradigm of an inductive and descriptive logic, being characterized as an exploratory and an ethnographic research (Angrosino, 2007; Coutinho, 2015; Gil, 1989; Lazar et al., 2017), within which we emphasize the correspondence of the researcher with the reality to be investigated, that is, the context of the PoC and its practices, where the construction of a theory is developed in an inductive and systematic way starting from its own natural habitat, as empirical data emerge (Creswell, 1994, 2012; Creswell & Creswell, 2018).

3.1. Objective 1: Identify, Classify, and Map the Connections in a PoC Context

The data collection resulted from our direct immersion in the natural PoC habitat. We totally immersed ourselves in the PoC world, over a period of approximately thirty (30) months, and analyzed how different IT companies, to a total of five (5), develop and execute PoC activities concerning several organizations (potential customers), based on the following criteria: (i) their ability to execute and develop the PoC activities, specifically in the domain of IT data infrastructures; (ii) their capacity with respect to the exposure and diversity in the development and execution of PoC activities, where we opted for companies that could offer diversity in the context of PoC activities, e.g., evaluating the performance of a data technology artifact; studying new features of high-availability scenarios, resilience, and data replication architectures; comparing different artifacts and technologies; among other scenarios; (iii) their ability to carry out these activities in parallel and not just as 'one or another eventual activity,' i.e., we opted for companies that could actually perform multiple and diverse PoC activities as their main operating factor; and (iv) in order to accept our observations and direct participation in PoC activities performed by those companies along with their practitioners.

We started our research as observers based on an ethnographic exercise (Lazar et al., 2017) in the natural PoC habitat, which according to Angrosino (2007, p. 14) is the "art and science of describing a human group — its institutions, interpersonal behaviors, material productions, and beliefs." In other words, this ethnographic exercise was based on the "notion that true understanding of complex human practices and contexts requires in-depth, engaged study" (Lazar et al., 2017, p. 231), in which we visualize PoC practitioners who often describe what they do in terms of movements and practices in the context of PoC, in a way that is actually inaccurate, possibly due to a lack

of awareness or understanding of what they would need to communicate about what they are aiming to accomplish (e.g., PoC outcome – results) (Neto et al., 2020a).

As observers, we were interested in identifying, characterizing, and tracking the concepts used by practitioners in the development and execution of PoC activity with the aim to contribute to our explicit reflections and interactive constructions regarding how those PoC elements (network nodes) evolve and relate in the flows (network connections) of information and knowledge. Therefore, we sought to incorporate our observations in a language that could be referenced and build the connections in the context of PoC. In this way, we observe what happens, listen to what is said, and ask specific questions through informal dialogues, thus recording this information through our personal notes and sketches (anonymized data) in a non-technical language (Angrosino, 2007; Hammersley & Atkinson, 2007; Lazar et al., 2017). In fact, we seek to gather whatever data are available to clarify the objectives that are the emerging focus of our investigation.

Notwithstanding, we opted to act based on two mindsets in the natural PoC habitat: non-interventionist (as observers) and interventionist (as practitioners). Thus, we worked collaboratively with diverse practitioners in the research domain, aiming to reflect on how the interactions (flows) between the PoC practices could contribute to knowledge management in the activity, in order to fulfill our objectives in this research. As both observer and practitioner, with the aim to accomplish the first objective in this research, our interest was in tracing (identifying and characterizing) how those concepts could help make explicit reflections, and our method was based on interactive observations and constructions on how those practices evolve and relate in information/knowledge flows across the PoC activity (Neto et al., 2020b), especially with respect to knowledge management in the context of PoC.

Later, we immersed ourselves again in the natural PoC habitat with a designer-researcher mindset (Hevner, 2007), aiming to develop our research based on interactive and practical constructions of how those nodes evolve and interact in that specific context (i.e., the movements used in PoC during a discussion of performance on a data storage system between an expert on PoC and a potential customer). In other words, when identifying a PoC practitioner or practice (node) being used in PoC, we observe its interactions (flows) with the next actions, i.e., competencies, movements, and scenarios, performed by PoC practitioners.

Thus, we created and analyzed an interactive design cycle (Hevner, 2007) within which we continually adapted a knowledge networking model through the next PoC activities, in order to discover how those practices, movements, competencies, and scenarios relate and evolve during execution, aiming to characterize a network in the context of PoC. That is, we contextualize the PoC activities during the translation phase when a PoC practitioner translates the phenomena of one world (e.g., organization world) to be reproduced in a different world (PoC world), and the "transition phase" (i.e., from the organization world to a PoC world) when a practitioner changes, combines, or improvises the phenomena in the PoC context between practices.

In summary, we aim to identify and characterize in those activities all the PoC connections in a manner relevant to interpreting and structuring action and the impact in the context, with respect to the production and dissemination of knowledge, with the aim to map the practices and their model elements in a knowledge network supported by the PoC practices context model. Also, we aimed to predict what are the next steps in modeling the context of PoC, especially regarding those connections, helping to perceive potential cluster nodes and their relations (connections) with the aim to identify and characterize the formation of the knowledge networks in the context of PoC.

3.2. Objective 2: A Proposal to "Measure" the Nodes Distributed in PoC Knowledge Networks

On the other hand, and with the aim to accomplish our second objective in this research, to measure those nodes distributed in the knowledge network, we needed to seek supporting evidence in a review of the literature. We found several and interesting studies in the field of physics (Sahimi, 1994; Stauffer & Aharony, 1992) related to Percolation Theory which describe the properties of those clusters (networks and sub-networks) where a percolation process consists of the propagation of the state from one vertex (node) to other vertices (nodes), which after being 'active' continue with the propagation process (percolation).

Thus, we envision this to be a way to understand whether this specific network is collaborating or not with the production or dissemination of knowledge within the practice in the context of PoC. In other words, the collaboration (activity) ends when there are no more vertices (nodes) in the network that can be activated, that is, when the network does not produce more knowledge. However, there are a number of centrality measures (Paula, 2015; Piraveenan et al., 2013) to determine the "importance" of

a single node (vertex) in a network (graph) or a subnet (sub-graph), but one potential approach to categorize and calculate the generated dynamic network is to use Percolation Centrality (Piraveenan et al., 2013, p. 3), which "measures the importance of nodes in terms of aiding the percolation through the network" and specifies in detail the importance of nodes (i.e., practices) in terms of their percolation (propagation) through the generated graph (i.e., PoC practices context model).

In order to accomplish our second objective, we seek the support of a Percolation Theory model, which we envisage to be useful for calculating a **learning curve** (our emphasis) for each knowledge network in the context of PoC, thus aiming to propose and address a dynamic percolation model in each network which could contribute to capturing the "relative impact of nodes during (possibly partial) percolation" (Piraveenan et al., 2013, p. 2) during PoC activities, and contribute to knowledge management in those activities. We found in the model proposed by Piraveenan et al. (2013) a measure of centrality (percolation centrality) with the aim to analyze the importance of nodes during percolation in networks, especially the relative impact of those nodes in additional percolation, e.g., in the spread of an infection, or in the context of our research, in the 'spread of knowledge' in PoC activities. That is, we envisage this percolation centrality model which becomes a useful measure and precisely in scenarios when an early intervention is needed, e.g., in order to relocate a practitioner from one practice to another or to strengthen a specific experiment with a practitioner who has greater 'knowledge percolation.'

Thus, we see percolation as an analogy to a "contagion" that occurs in complex networks in several scenarios. For example, in the case of a viral infection, it can spread through people's social networks. Likewise, the spread of diseases can also be considered as a higher level of abstraction, covering several networks, e.g., a network of cities connected by roads. In the same way, computer viruses can spread over computer networks and the whole Internet. Therefore, in all of these scenarios, somehow a "contagion" happens and spreads over the links of a complex network, changing the 'states' of the nodes as it spreads.

In the context of our research, we envisage the use of Percolation Theory models with the aim to contribute to the development of a proposal, albeit theoretical, of how communities of practice could identify, characterize, and measure this "contagion" in the PoC Knowledge Networks, where this "contagion" refers to the information and knowledge spread across all the connections of those networks which somehow impacts the production and dissemination of knowledge.

4. RESULTS AND DISCUSSION

Our research was based on our direct participation (Angrosino, 2007; Lazar et al., 2017) in multiple PoC activities, totaling eighty (80) – fifty (50) in total as non-interventionists (observers), and thirty (30) in total as interventionists (practitioner) – working collaboratively with several PoC practitioners, ninety-seven (97) in total. As a result, we observed ten (10) practices, twenty-two (22) competencies, one hundred and ten (110) movements (action-codes) and four hundred and forty-three (443) data points, specifically three hundred and ninety-one (391) PoC scenarios (e.g., dialogues) and fifty-two (52) sketches in the context of PoC. After mapping all the potential interactions (connections) between the practices, competencies, movements (action-codes), and scenarios (Neto et al., 2019) in the context of PoC, we identified a total of five hundred and eighty-five (585) nodes and one thousand and sixty-nine (1,069) connections (links) among all the practices, competencies, movements, and scenarios based on the PoC practices context model (Neto et al., 2020b), composing a large and complex knowledge network (Table 1).

Phelps et al. (2012) describe networks as "a set of nodes — individuals or higher level collectives that serve as heterogeneously distributed repositories of knowledge and agents that search for, transmit, and create knowledge — interconnected by social relationships that enable and constrain nodes' efforts to acquire, transfer, and create knowledge" (Phelps et al., 2012, p. 1117), i.e., we envisage these PoC networks as **knowledge networks** (our emphasis).

In other words, Davenport and Prusak (1998, p. 116) note "there are 'knowledge networks' for each key practice." Thus, each connection (link) in those networks demonstrates a **higher probability for the production and dissemination of knowledge** (our emphasis) through the connected nodes which are represented by the practices, the competencies, the movements, and the scenarios, with a potential influence over the whole network.

Further, with the aim to demonstrate some of the knowledge networks found in the context of PoC, we illustrate several networks (Fig. 4) based on the practice of Exploring (Neto et al., 2020b, 2019) composed of one (1) practice, seven (7) competencies, eighty-five (85) movements, and three hundred and twenty-five (325) PoC

Table 1. Nodes and links (connections) in the context of Proof-of-Concept

Practices	Nodes	Links
Practice of Exploring	418	523
Practice of Comprehending	358	443
Practice of Modeling	446	565
Practice of Specifying	477	628
Practice of Executing	540	836
Practice of Negotiating	505	737
Practice of Improvising	488	761
Practice of Reflecting	527	751
Practice of Describing	425	550
Practice of Documenting	490	640

scenarios, totaling four hundred and eighteen (418) nodes and five hundred and twenty-three (523) connections (links).

As can be seen in Fig. 4, Network 1 represents a network constituted by 8 nodes (one practice and seven competencies). Network 2 represents the 7 competencies and their respective movements (action-codes), and Network 3 represents the movements (action-codes) and PoC scenarios in the practice of Exploring.

Thus, we envisioned and learned about these nodes and their connections (links) creating and promoting knowledge; and knowledge that will reproduce other nodes and connections, thus reifying and evolving knowledge of the practice in PoC activities, where "[o]ne, knowledge is a function of a particular stance, perspective, or intention. Second, knowledge, unlike information, is about action. It is always knowledge 'to some end.' And third, knowledge, like information, is about meaning. It is context-specific and relational" (Nonaka & Takeuchi, 1995, p. 72).

4.1. The Production of Meaning in PoC Knowledge Networks

However, during our observations and participation in the natural PoC habitat, we discovered an interesting phenomenon in regard to the production of discourse and knowledge in the context of PoC being based on the relationship of the parts with the whole and vice-versa (that is, think here of the nodes and their connections in a network) or the mediators with their context, that is, elements that allow the production of meaning (Roque, 2004).

In this way, we understand that reality capture is performed when the practitioner in the context of PoC really seeks to perceive the world in which he/she lives, not only based on his/her *sui generis* perspective, but also in relation to the different perspectives and actions of other actors in the context of this activity, where we emphasize that this set of different perspectives in the context of PoC enhances, and at the same time provokes, new thoughts and reflections in order to contribute to the formation and improvement of knowledge networks in the context of PoC.

In other words, the production of meaning is grounded by the hermeneutic circle (Gadamer, 2008, 2013; Neto et al., 2020a; Nixon, 2017; Schmidt, 2012) and based on the movements of interpretation and comprehension using the knowledge networks in the context of PoC, where during these movements of interpretation and understanding, practitioners and their activities create and promote knowledge, where that knowledge contributes to their own evolution, in addition to the production of new activities (Fig. 5).

As we can see in Fig. 5, we present one scenario observed during our immersion in the PoC world. The scenario was composed of two practitioners working in the context of PoC supported by the formation of cycles of dialectical reproduction of their activities, oriented towards the action of a body of knowledge based on the interpretation and understanding of all parties (or the knowledge network nodes) in the context of PoC. Note that during our observations and participation in the natural PoC habitat, we identified several passages in the scenarios in the PoC that supported our understanding regarding the hermeneutic nature of this activity.

In one of these observations, we highlight a particular situation where a customer would like to use a synthetic workload tool for simulating I/O (Input/Output) operations with a very large number of threads (processes) in order to evaluate a new technology, which in our opinion is somewhat absurd. Upon receiving this request, the group of PoC practitioners interpreted it as unrealistic and based on "their knowledge" (production of meaning), it is unnecessary to use an amount of 65,536 threads in a performance evaluation (I/O and latency) experiment of a data storage system (Almeida, 2006; Neto, 2004).

However, from the customer's point of view, we noticed an impasse, where "after numerous discussions between the solutions architects and the PoC group, the customer was reluctant to change the proposed experiments, especially the specific experiment related to the 65,536

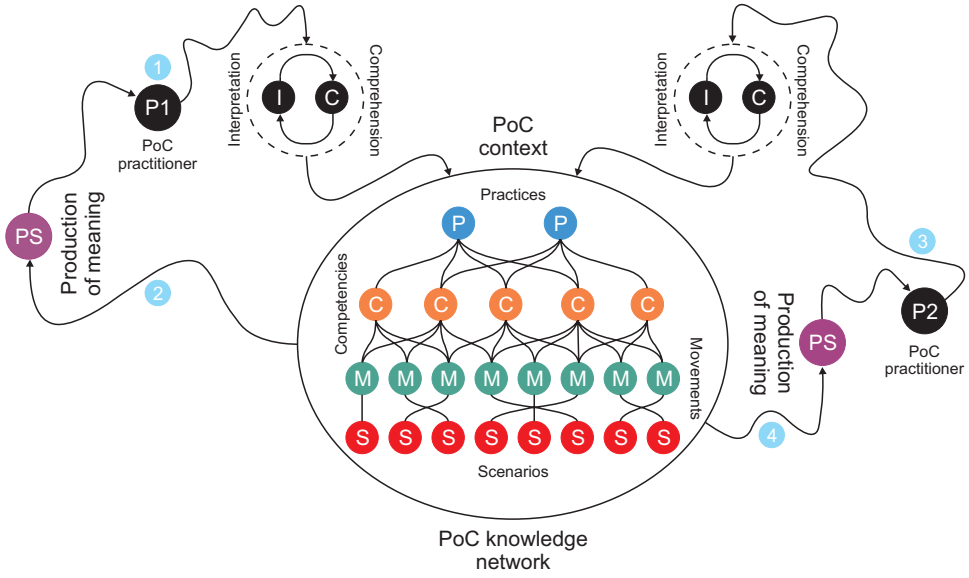


Fig. 5. The production of meaning in the context of Proof-of-Concept (PoC).

threads" (Neto et al., 2020a, p. 129).

Thus, this specific experiment was rejected by the group of PoC practitioners, who initially understood that this number of threads (processes in parallel) characterized a misconfiguration (that is, the non-representation) of any information systems in reality.

In other words, the interpretation/comprehension (hermeneutic cycle) of this group was as follows:

> ...it is not necessary to use this number of processes in parallel [65,536 threads] for the generation and simulation of the data for performance measurement in data storage systems. Also, the results obtained could express an outcome that could be misinterpreted, thus formalizing a deficient knowledge based on these results (Neto et al., 2020a, p. 129).

In the course of countless discussions between the technical staff of the customer, the business partner, and the group PoC practitioners regarding this specific experiment (65,536 threads), we identified several recurring cycles of interpretation and understanding for the production of meaning (again, the hermeneutic cycle). Thus, we understand that in order for this production of meaning to make "some sense," an engineering of a context becomes necessary. That is, we emphasize that these cycles of interpretation and understanding, the production of meaning, and the engineering of a context, that is, a hermeneutic circle associated with the PoC context, happened naturally in the course of the countless conversations between these actors.

At the end, after several discussions and recurring cycles of interpretation and understanding in the context of the PoC (Fig. 5), the group of PoC practitioners realized that the customer's intention (i.e., the specific objective) with this particular experiment **was not to evaluate the performance** (our emphasis) of the data storage solution, but to understand and explore how this technology would behave under an abnormal condition of use (i.e., how this technology would behave 'dealing' with 65,536 processes in parallel). The customer explained that, in the past, they had this experience with another solution from a different manufacturer, and when that situation happened due to a misconfiguration application, the solution was virtually unusable, consequently impacting their core business. So that would be why they would like to learn and understand how this new data storage solution would behave under the same conditions.

As can be seen in Fig. 5, we learned that every understanding in the PoC context is born with prejudices (preconcepts), where practitioners adopt, modify, or reject these prejudices as the development of this activity continues. Thus, we envisage this development proceeding in a hermeneutic circle (Gadamer, 2013) (see the different circles of interpretation and comprehension in Fig. 5) grounded by the formation of knowledge networks in the context of PoC.

In other words, we define a movement, which is precisely a back-and-forth movement between those prejudices that lead to practices in the context of PoC and the knowledge that we derive from that movement. Thus, if there were no prejudices to modify, understanding would

be impossible and unnecessary. However, the practitioner in the PoC context is not an isolated node in the knowledge network, where that practitioner and the knowledge live in a given context (network). Both interact with this context (network); at the same time, they influence and are influenced by the context (those networks) (Neto et al., 2020c).

Therefore, the hermeneutics presents

> itself in the present time, in the culture of a particular group, to seek the meaning that comes from the past or even the present, from a worldview of its own, involving in a single movement the being who understands and also what is understood, which is specifically what happens in the PoC world (Neto et al., 2020a, p. 130).

Thus, reflecting on what happens in that world supported by hermeneutics can reveal problems and help clarify and understand the dimension of the production of meaning or the phenomena inherent in the context model of PoC practices (Neto et al., 2020c), especially with the formation of those knowledge networks.

4.2. Connecting PoC Knowledge Networks and Percolation Theory in the Context of PoC

In the same sense, in the context of PoC we learned that the production and dissemination of knowledge are dynamic through social interactions (Nonaka et al., 2000) among its practitioners across the entire network. Thus, we envisioned flows of information/knowledge also when PoC practitioners (acting as nodes in a network) "spread" from one to another implicit and explicit (tacit) knowledge throughout PoC Knowledge Networks. In other words, using an analogy as an example, we can represent these flows of information/knowledge in PoC Knowledge Networks as a "contagion" of a virus on a computer network or a dissemination (spread) of information in a social network.

As can be seen, PoC through its practitioners (players) produces results and these results (Neto et al., 2019, 2020a, 2020b), such as new scenarios, new movements, new competencies, new practices, new knowledge, even new practitioners, or any element could be represented as new nodes (vertices) to be used collaboratively in the network itself, parts of the core network, or in multiple networks, which can be represented through graphs (Feofiloff et al., 2011; Gonçalves, 2007; Jurkiewicz, 2009; Lucchesi, 1979) as a mathematical structure that is used to represent the relationships between clusters (objects).

Thus, using the terminology of Graph Theory (Feofiloff et al., 2011; Lucchesi, 1979), it can be said that the distribution of *P(k)* connectivity that gives the probability that one vertex (e.g., practitioner, practice, or knowledge) is connected to *k* other vertices, is mathematically described by a power-law $P(k) \alpha k^{-\gamma}$, where the factor γ will depend on the scale of the considered network.

The network growth begins with random graphs of *k0* vertices, and each vertex *k* could be considered as just a simple new practitioner, new movement, new practice, or knowledge generated in the context of PoC. The growth hypothesis is incorporated assuming that in each unit of time *t* a new vertex is added to the graph, connecting to *k* different vertices. At each time point *t* each vertex *k* has a probability *p* of acquiring a new connection, that is, knowledge was generated or propagated to *k* other vertices, or a new PoC practitioner had connected to the network, with the objective of establishing new alliances and/or consuming/producing knowledge.

According to Piraveenan et al. (2013), the percolation state of a node (vertex) *k* at time *t* is given by x_k^t. Specifically, $x_k^t=0$ indicates a non-percolated state at time *t*; where $x_k^t=1$, a fully percolated state; or we can say that $0<x_k^t<1$ would be a partially percolated state at time *t* (Fig. 6).

Note that we use hypothetical values to demonstrate the percolation states of each node, which could be determined (and this is to be explored in future works) as, for example, the frequency of PoC movements for each practitioner or the amount of evidence (the total circulation inside a particular network), such as the number of documents, results, consultations, publications, or participations in PoC activities, among others. We undoubtedly recognize that more research is needed in this particular area regarding the percolation states in these networks, i.e., it is essential to define, depending on particular PoC practices, how to evaluate each node in the PoC network or the node percolation states in a PoC Knowledge Network, among other scenarios. Also, we recognize the need for further study in order to analyze the 'culture' of knowledge sharing in the context of PoC due to its heterogeneity and complexity. It is known that the type of knowledge in the context of PoC practices implicitly or tacitly influences the route choice and speed of knowledge sharing. It would be interesting to investigate further what types of knowledge and what practices affect the percolation threshold.

Thus, as can be seen in Fig. 6, this particular network (the subnet inside the dotted lines) has five vertices (nodes in green), where each vertex (node) has its respective

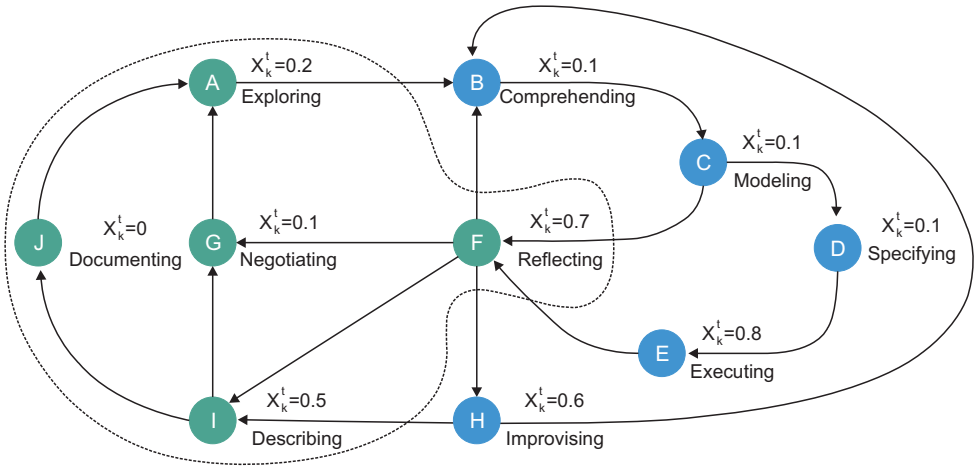


Fig. 6. An example of vertices (nodes) and Percolation states (values) in a Proof-of-Concept Knowledge Network.

state (weight) of percolation as: **Vertex F** has a percolation status (weight) of x_k^t=0.7; **Vertex I** has a percolation status (weight) of x_k^t=0.5; **Vertex J** has a percolation status (weight) of x_k^t=0; **Vertex G** has a percolation status (weight) of x_k^t=0.1; and **Vertex A** has a percolation status (weight) of x_k^t=0.2.

However, Vertex F (x_k^t=0.7) has a percolation status (weight) greater than vertex I (x_k^t=0.5), whereby we could conclude that the information or knowledge generated from that vertex or existing in it has a "higher probability to percolate," that is, propagate to vertex I. Likewise, we could conclude the same with respect to vertex I (x_k^t=0.5) in relation to both vertices J (x_k^t=0) and G (x_k^t=0.1).

In the case of vertex J (x_k^t=0), we could conclude that nothing percolates, that is, nothing passes through this vertex, since it has a zero percolation state (weight). In the case of Vertex G (x_k^t=0.1), in relation to vertex A (x_k^t =0.2), we can conclude that the "probability of percolating is low," because the percolation state (weight) of this vertex is lower in relation to the subsequent vertex (Vertex A).

Thus, the main purpose of Percolation Centrality, according to Piraveenan et al. (2013, p. 3), is based on the fact that "a node may very well be centrally located in terms of betweenness centrality or another centrality measure, but may not be 'centrally' located in the context of a network in which there is percolation," e.g., two practitioners (players 1 and 2) and their impact on other practitioners (vertices) in the entire network (graph). According to the same authors, there is a series of centrality measures that "exists to determine the 'importance' of a single node in a complex network. However, these measures quantify the importance of a node in purely topological terms, and the value of the node does not depend on the 'state' of the node in any way" (Piraveenan et al., 2013, p. 3).

As can be seen in Fig. 3, there are several clusters, that is, several groups of players (i.e., PoC practitioners) and their vertices (e.g., practices). This potential network is based on Percolation Centrality (Piraveenan et al., 2013) and focuses on determining **how important** (our emphasis) (regarding quality) the node (player or knowledge) is in the complete percolation process, which corresponds to the "spread of knowledge to other networks," i.e., PoC Knowledge Networks. The Percolation Centrality proposed by Piraveenan et al. (2013, p. 4) of a given node (vertex) *k* at time *t* can be represented by (1):

$$PC^t(k) = \frac{1}{(N-2)} \sum_{s \neq k \neq r} \frac{\alpha_{s,r}(k)}{\alpha_{s,r}} \frac{x_s^t}{[\sum x_i^t] - x_k^t} \quad (1)$$

where *PC* is Percolation Centrality, which in the context of this research could indicate the value of the spread of knowledge at each node (player or knowledge); *k* represents a node; $\alpha_{s,r}$ represents the number of shortest paths between the source node *s* and the destination node *r*; x_k^t corresponds to the node percolation state at a given time *t*; and Σx_i^t represents the sum of the node percolation state represented by the variable *x* at a given time *t*. In other words, the higher the $PC^t(k)$, the greater the degree of knowledge spread of that node (e.g., PoC practitioner). The weight (state) linked to the percolation paths depends on the percolation levels signed for the originating nodes, based on the premise that the higher the percolation level of an originating node, the more important the originating paths of that node are.

For instance, taking Fig. 3 as an example, the dissemination of knowledge produced in the practice of Improvising (Network Y) to the practice of Describing could be questioned, if the source and destination nodes have equal

levels of percolation (propagation), where the paths that connect them become insignificant for a possible transmission of knowledge. However, if the destination nodes have lower percolation levels, the knowledge produced by the practice of Improvising tends to percolate (propagate), i.e., the dissemination of knowledge is only disseminated if the nodes of origin are at higher levels compared to the destination nodes.

4.3. A Conceptual Example of Using Percolation Theory and Knowledge Networks in PoC

In the following, we aim to demonstrate two scenarios using a PoC practices context model (Neto et al., 2020b) supported by Knowledge Networks and Percolation Theory, particularly the Percolation Centrality model proposed by Piraveenan et al. (2013).

In both examples, during our immersion in the natural PoC habitat, we observed PoC practitioners forming knowledge networks, but in different practices (Fig. 7). Note that the percolation thresholds used in the percolation probability in both examples in this section are intended to be conceptual illustrations. That is, if we visualize a node (e.g., a player or practice) with a value of 0.2 (percolation state), it has higher probability to percolate (propagate) knowledge when compared to a node with a value of 0.1.

As can be seen in Fig. 7, we have 3 players (PoC practitioners) – Player 2, Player 4, and Player 6 generating a certain knowledge – Knowledge A, which has a probability of percolation=0.2 in the practice of Exploring with the following probabilities of percolation: (i) Player 2, probability of percolation=0.1; (ii) Player 4, probability of percolation=0.2; and (iii) Player 6, probability of percolation=0.2.

On the other side, the knowledge – Knowledge H generated by Players 9 and 7 has the following probabilities of percolation: (iv) Player 9, probability of percolation=0.5; and (v) Player 7, probability of percolation=0.5 in the practice of Improvising.

Considering that node Knowledge A has the largest number of players and connections, **we could deduce** (our emphasis) that the centrality of percolation would be higher than node Knowledge H, thus, having a higher influence on node Knowledge B.

However, calculating the Percolation Centrality based on Fig. 8, the PC of node Knowledge A, represented as *PC (ka)*, is 0.625 and the PC of node Knowledge H, represented as *PC (kh)*, is 0.667. In other words, even if node Knowledge A has a higher number of players and connections, the higher probability of percolation and potentially more influence on node Knowledge B is with node Knowledge H, which has a higher level of PC because it is connected to players with a higher level high of PC (Fig. 8).

In a different experiment, we observed and learned from our data analysis that if the PoC Knowledge Network has a different formation (Fig. 9), it can contribute with a higher or lower probability of percolation, thus influencing other nodes in the network. In other words, a different network formation with other players or players in different positions (i.e., different practices) can affect (positively or negatively) the production and dissemination of knowledge throughout the whole network.

As can be seen in Fig. 9, this PoC Knowledge Network has a different formation compared to Fig. 7. In this example, we have the players – Players 1, 9, and 7 generating some certain knowledge – Knowledge A with a probability of percolation=0.2 in the practice of Exploring with

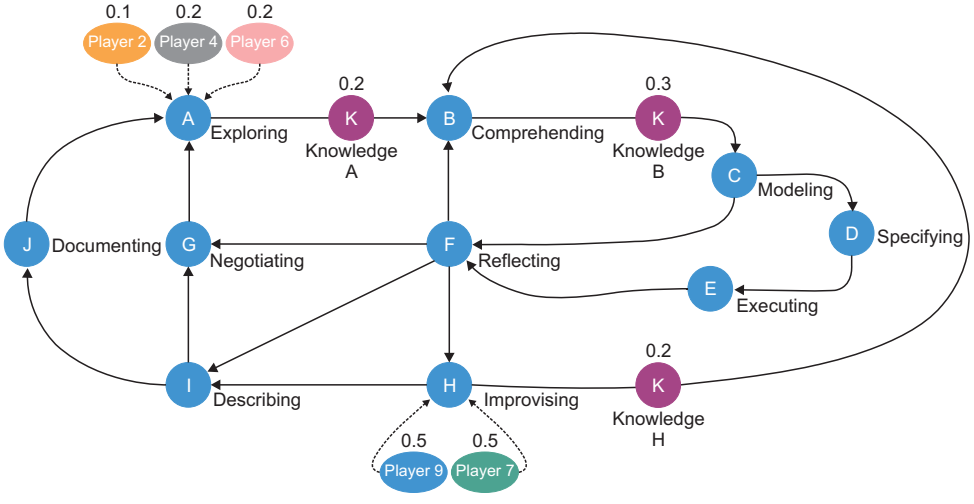


Fig. 7. Proof-of-Concept Knowledge Network – nodes and their probability (values) of percolation.

 https://doi.org/10.1633/JISTaP.2021.9.1.1

the following probabilities of percolation: (i) Player 1, probability of percolation=0.3; (ii) Player 9, probability of percolation=0.5; and (iii) Player 7, probability of percolation=0.5.

On the other hand, the knowledge – Knowledge H generated by Players 4 and 5 has the following probabilities of percolation: (iv) Player 4, probability of percolation=0.1; and (v) Player 5, probability of percolation=0.1 in the practice of Improvising. Thus, calculating Percolation Centrality based on Fig. 9, the *PC* of node Knowledge A *(ka)* is 0.825 and the *PC* of node Knowledge H *(kh)* is 0.4. We understand that node Knowledge A has a higher level of *PC* as it is closer to players (PoC practitioners) with a higher percolation value and a greater probability to influence node Knowledge B (Fig. 10).

If we consider, in the case of Fig. 8, the addition of a new player (PoC practitioner) to the 'game' (e.g., PoC activity), there is a high probability of the establishment of a new network (alliance) of the new player towards another player, which increases the random acquisition of a new knowledge (new node) using node Knowledge H, which has a PC greater (0.677) than the PC of node Knowledge A (0.625), in this particular practice.

However, in the case of Fig. 10, the probability of acquiring knowledge increases in the use of node Knowledge A, which has a PC greater (0.825) than the PC of node Knowledge H (0.4), in this particular practice.

4.4. The Applicability of Knowledge Networks and Percolation Theory in the Context of PoC

While PoC is an activity that appears to be "purely technological," particularly in the domain of IT, its context is grounded in a complex network (or several networks) of sociotechnical phenomena (Neto et al., 2019). In other words, there are various "free and intuitive" movements carried out by its set of practitioners forming the core of

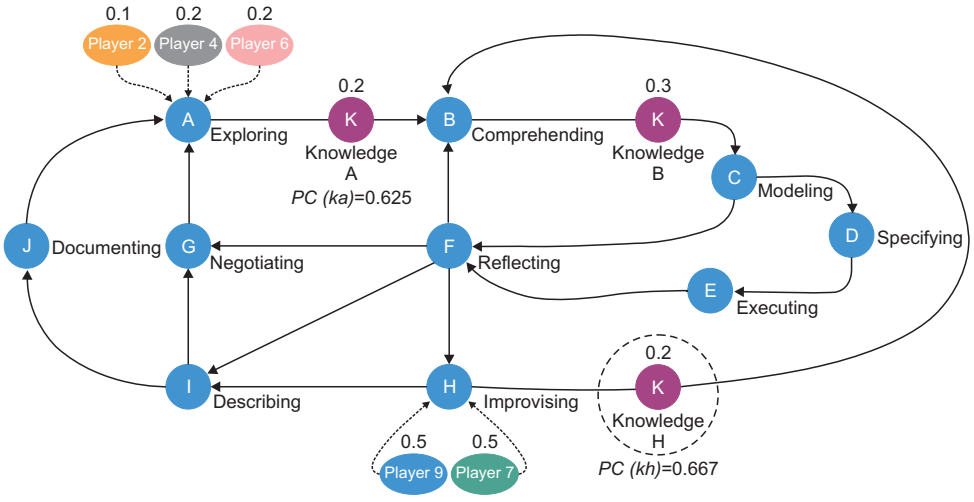


Fig. 8. Node Knowledge H has a higher *PC (kh)* – a higher probability to influence node Knowledge B.

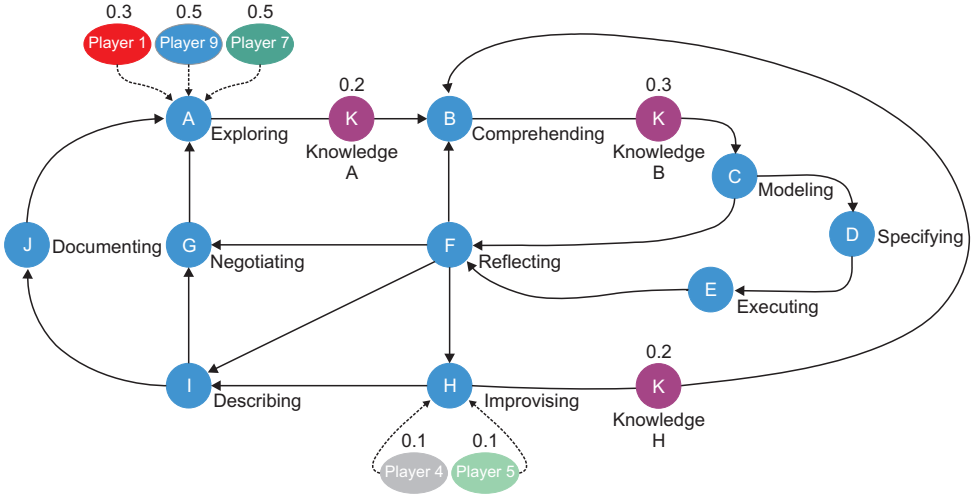


Fig. 9. Proof-of-Concept Knowledge Network – different players in the practice of Exploring and Improvising.

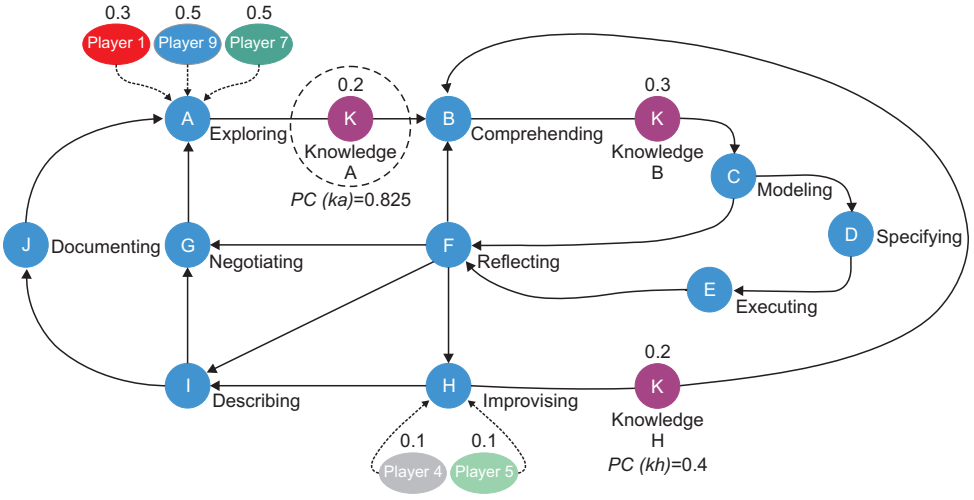


Fig. 10. Knowledge A has a higher *PC (ka)* – a higher probability of influencing Knowledge B.

these practices in the context of PoC, acting without a rationality and explicit catalog of activities and in constant cycles of collaboration with different actors (practitioners and several organizations) during development and execution of PoC.

Therefore, we learned that in the context of PoC it is **not always what it looks like or is imagined to be** (our emphasis), but rather "in theory there is no difference between theory and practice, while in practice there is" (Brewster, 1882, p. 202).

In this way, we envision that in the use of Practice Context Models 'linked' to Knowledge Networks and Percolation Centrality, there is a contribution to an analysis of the taxonomy among players (e.g., PoC practitioners, specialists, enthusiasts, researchers, teachers, etc.), certain knowledge (produced or consumed), and their research networks. With the adoption of this proposal each player has a model in the creation of alliances (knowledge networks) with different players (different PoC practitioners), in order to fulfill the 'game mission' (e.g., generation of scientific and technological knowledge in the context of PoC).

In other words, we envisage the use of "Practice Context Models linked to Knowledge Networks and Percolation Theory" as a potential (and feasible) proposal to be built through the attribution of values (weights) to the vertices (players, practices, competencies, movements, scenarios, and knowledge). This approach can allow the players (PoC practitioners) to have greater flexibility in building alliances with other players (new vertices), focusing on those with higher value (focus on quality) in collaboration networks (research networks). In other words, alliances (connections) aim to produce and make available knowledge in complex networks in the context of PoC, thus contributing to knowledge management in the context of PoC.

Hence, practitioners and organizations in the context of PoC could use this proposal to contribute to the definition of an evaluation model within practices in the context of PoC. In this way, the Practices Context Model could contribute to the analysis and documentation of the formation of those knowledge networks in PoC and with that establish their states, based on several and specific criteria and an evaluation defined by any organization, and with this they could contribute to the improvement of those practices in the context of the PoC, as well as the whole 'PoC practice.'

5. CONCLUSION AND FUTURE WORK

According to Borges (2006), the universe of scientific information is a universe of sharing – ideas, techniques, methodologies, **information**, and **knowledge** (our emphasis), a definition that we translated to and apply to a PoC practices context model. Therefore, we highlight a definition of a collaborative, interactive, and motivating model in the establishment and connection of different PoC practices between different practitioners and their organizations, which aims to develop and provide an innovative collaboration platform in a PoC context.

During our immersion in the natural PoC habitat, we discovered an interesting phenomenon with respect to the production of discourse and knowledge in the context of PoC being based on the relationship of the parts with the whole and vice versa or the mediators with their context, that is, elements that allow the production of meaning. In other words, we learned the production of meaning is grounded by the hermeneutic circle and based on the

movements of interpretation and comprehension using the knowledge networks in the context of PoC, where during those movements of interpretation and understanding, practitioners and their activities create and promote knowledge, where that knowledge contributes to their own evolution, in addition to the production of new activities.

Also, we identify the development of several knowledge networks in the context of PoC, e.g., a network where two practitioners work together, through the practice of Comprehending, in search of more information about a particular artifact to be used in PoC (Fig. 3). Another example is a network composed of a set of practitioners seeking alternatives (Practice of Improvising) to replace a performance benchmark that was initially considered to be used in a PoC. Thus, we envisage PoC practitioners using the Practices Context Model associated with Knowledge Networks and Percolation Theory, aiming to encourage collaboration between different practitioners and promoting the construction of new knowledge networks in the context of PoC.

In other words, we see a greater likelihood of the development and improvement of PoC practices in order to contribute to the production and dissemination of knowledge, through a naturally collaborative and user-friendly platform, aiming to contribute to the systematization and evaluation of these networks. Moreover, the logic of the model could intelligently distribute the participating nodes (e.g., PoC practitioners) in their missions (PoC practices) in order to contribute so that the entire network can percolate (propagate) information and generate knowledge.

In this study, we have learned about the 'natural knowledge' in PoC that falls on a **spectrum of cognitive development spread throughout its network** (our emphasis), as each element (node) in the network produces and disseminates certain knowledge that flows to other elements (e.g., PoC practitioners, practices, other PoC activities, among others; see Fig. 3). Therefore, we present this study as a starting point of novel research into using Practice Context Models with the aim to contribute to knowledge management in PoC activities. We envision the "connection" with Knowledge Networks and Percolation Theory as a potential instrument to identify the network structure based on a percolation threshold and analyze the importance of the nodes (PoC practitioners, practices, competencies, movements, and PoC scenarios) during the percolation of knowledge in these networks in the context of PoC.

As the main objective in this study, we aimed to provide scenarios (Neto et al., 2019, 2020a, 2020b) grounded by a PoC practices context model of how the practitioners, practices, competencies, movements, and scenarios in the context of PoC could learn and then influence (and be influenced by) the percolation threshold. We understand that this would be useful for calculating a learning curve for each knowledge network in aiming to address a dynamic percolation model in each practice in the context of the PoC model, where a PoC practitioner can excel in one particular practice (e.g., Practice of Improvising), but be deficient in another practice (e.g., Practice of Documenting).

For future work, we intend to materialize this model in a software development prototype and propose methods on how to evaluate and classify the nodes in those PoC Knowledge Networks. Thus, we aim to advance our research in seeking possibilities regarding how to make dynamic predictions about future PoC developments and movements by their practitioners, i.e., how to distribute the players (PoC practitioners) in the knowledge networks, for example, players 1, 2, and 3 (PoC practitioners), aiming to provide a balance across the whole network and contribute to knowledge management over the entire network (practice) in the context of PoC.

CONFLICTS OF INTEREST

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

REFERENCES

Almeida, A. V. (2006). *Arquitetura de redes de armazenamento de dados* [Architecture of data storage networks] [Master's thesis, Universidade Estadual de Campinas, Campinas, Brazil]. http://repositorio.unicamp.br/jspui/handle/REPOSIP/276100

Angrosino, M. (2007). *Doing ethnographic and observational research*. Sage.

Barabási, A.-L. (2016). *Network science*. Cambridge University Press.

Barnes, J. D., Katzer, R. D., Potluri, D., & Stone, M. J. (2009). Demonstrating proof of concept of a project with requirements component providing weights on importance and tracking with use cases (U.S. Patent No. 7,509,626). U.S. Patent and Trademark Office. http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=7509626.PN.&OS=PN/7509626&RS=PN/7509626.

Block, M., Khvatova, T., Zhukov, D., & Lesko, S. (2015, November 12-13). Studying the structural topology of the knowledge sharing network. In J. C. D. Rouco (Ed.), *Proceedings of the 11th European Conference on Management Leadership and Governance* (pp. 20-27). Academic Conferences and Publishing International Limited.

Borges, M. M. (2006). *A esfera: Comunicação académica e novos media* [The sphere: Academic communication and new media] [Doctoral dissertation, Universidade de Coimbra, Coimbra, Portugal]. http://estudogeral.sib.uc.pt/handle/10316/8557

Brewster, B. (1882). Portfolio: Theory and practice. *Yale Literature Magazine*, 47(5), 201-205.

Coutinho, C. P. (2015). *Metodologia de investigação em ciências sociais e humanas: Teoria e prática* [Research methodology in social and human sciences: Theory and practice] (2nd ed.). Almedina.

Creswell, J. (2012). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Sage.

Creswell, J. W. (1994). *Research design: Qualitative & quantitative approaches*. Sage.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business School Press.

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Orienta-Konsultit.

Engeström, Y. (2001). *Expansive learning at work: Toward an activity theoretical reconceptualization. Journal of Education and Work*, 14(1), 133-156. https://doi.org/10.1080/13639080123238

Engeström, Y. (2016). *Studies in expansive learning: Learning what is not yet there.* Cambridge University Press. https://doi.org/10.1017/CBO9781316225363

Fanfan, Y. (2012, October 20-21). Knowledge management: Making a core competency in today's business world. In Institute of Electrical and Electronics Engineers (IEEE) (Ed.), *2012 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization* (pp. 118-121). IEEE. https://doi.org/10.1109/ICSSEM.2012.6340781

Feofiloff, P., Kohayakawa, Y., & Wakabayashi, Y. (2011). *Uma introdução sucinta à teoria dos grafos* [A brief introduction to graph theory]. https://www.ime.usp.br/~pf/teoriadosgrafos/texto/TeoriaDosGrafos.pdf

Fino, C. N. (2001). *Vygotsky e a Zona de Desenvolvimento Proximal (ZDP): Três implicações pedagógicas* [Vygotsky and the Zone of Proximal Development (ZDP): Three pedagogical implications]. *Revista Portuguesa de Educação,* 14(2), 273-291. https://digituma.uma.pt/handle/10400.13/799

Gadamer, H.-G. (2008). *Philosophical hermeneutics.* University of California Press.

Gadamer, H.-G. (2013). *Truth and method*. Bloomsbury Academic.

Gil, A. C. (1989). *Métodos e técnicas de pesquisa social* [Methods and techniques of social research] (2nd ed.). Atlas.

Gonçalves, A. L. (2007). *Grafos: Aplicações ao jogo* [Graphs: Applications to the game] [Master's thesis, Universidade Portucalense, Porto, Portugal]. http://repositorio.uportu.pt/xmlui/handle/11328/539

Hammersley, M., & Atkinson, P. (2007). Ethnography: Principles in practice. (3rd ed.). Routledge.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2), 87-92. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1017&context=sjis

Isabelle, B., Cécile, G., Carole, D.-G., Pascal, L., Jean, N., & François, P. (2012). *Coordination practices in extreme situations. European Management Journal*, 30(6), 475-489. https://doi.org/10.1016/j.emj.2012.03.015

Jacobetty, P. (2010). *Ciência aberta: Produção de conhecimento científico na sociedade em rede* [Open science: Production of scientific knowledge in the network society] [Master's thesis, ISCTE – Instituto Universitário de Lisboa, Lisboa, Portugal]. https://repositorio.iscte-iul.pt/handle/10071/3024

Jurkiewicz, S. (2009). *Grafos – Uma introdução* [Graphs – An introduction]. Olimpíada Brasileira de Matemática Das Escolas Públicas. http://147.65.23.4/docs/apostila5.pdf

Kendig, C. E. (2016). What is proof of concept research and how does it generate epistemic and ethical categories for future scientific practice? *Science and Engineering Ethics*, 22(3), 735-753. https://doi.org/10.1007/s11948-015-9654-0

Klaus, V., Schmidt, S., Schreiber, D., & Bessi, V. G. (2016). *Criação de conhecimento em uma empresa de software* [Knowledge creation in a software company]. Pensamento & Realidade, 31(3), 1-18.

Kuramoto, H. (2006). *Informação científica: proposta de um novo modelo para o Brasil* [Scientific information: Proposal for a new model for Brazil]. *Ciência Da Informação,* 35(2), 91-102. https://doi.org/10.1590/S0100-19652006000200010

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction.* (2nd ed.). Elsevier.

Lucchesi, C. L. (1979). *Introdução à teoria dos grafos* [In-

troduction to graph theory]. Instituto de Matemática Pura e Aplicada, Brazil. https://impa.br/wp-content/uploads/2017/04/12_CBM_79_05.pdf

Menczer, F., Fortunato, S., & Davis, C. A. (2020). *A first course in network science*. Cambridge University Press.

Neto, A. J. R. (2004). *Um estudo do desempenho dos protocolos iSCSI e fibre channel* [A study of the performance of the iSCSI and fiber channel protocols] [Master's thesis, Universidade Estadual de Campinas, Campinas, Brazil]. http://repositorio.unicamp.br/jspui/handle/REPOSIP/276309

Neto, A. J. R., Borges, M. M., & Roque, L. (2018, October 24-26). A preliminary study of proof of concept practices and their connection with information systems and information science. In F. J. García-Peñalvo (Ed.), *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 270-275). ACM. https://doi.org/10.1145/3284179.3284226

Neto, A. J. R., Borges, M. M., & Roque, L. (2019, August 28-30). Characterizing proof-of-concept practices using the lens of context engineering. In A. Siarheyeva, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), *Information Systems Development: Information Systems Beyond 2020 (ISD2019 Proceedings)* (pp. 1-12). ISEN Yncréa Méditerranée. https://aisel.aisnet.org/isd2014/proceedings2019/ISDMethodologies/4

Neto, A. J. R., Borges, M. M., & Roque, L. (2020a). A "new" view of proof-of-concept practices through the lenses of activity theory and hermeneutics. *Journal of Information Technology Research,* 13(4), 118-135. https://doi.org/10.4018/JITR.2020100108

Neto, A. J. R., Borges, M. M., & Roque, L. (2020b, June 15-17). Developing a proof-of-concept practices context model. *Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference* (pp. 1-14). https://aisel.aisnet.org/ecis2020_rp/174

Neto, A. J. R., Borges, M. M., & Roque, L. (2020c). Flow of knowledge in proof-of-concept activities: Examining the problem of interpretation using hermeneutics. *LIBRES*, 30(2), 45-67. https://www.libres-ejournal.info/3254

Nixon, J. (2017). *Hans-Georg Gadamer: The hermeneutical imagination.* Springer. https://doi.org/10.1007/978-3-319-52117-6

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How the Japanese companies create the dynamics of innovation*. Oxford University Press.

Nonaka, I., Toyama, R., & Konno, N. (2000). SECI, *Ba* and leadership: A unified model of dynamic knowledge creation. *Long Range Planning*, 33(1), 5-34. https://doi.org/10.1016/S0024-6301(99)00115-6

Paula, S. M. C. (2015). *Modelos elementares de percolação* [Elementary percolation models] [Master's thesis, Universidade Federal de Goias, Goiânia,Brazil]. http://repositorio.bc.ufg.br/tede/handle/tede/4802

Phelps, C., Heidl, R., & Wadhwa, A. (2012). Knowledge, networks, and knowledge networks: A review and research agenda. *Journal of Management*, 38(4), 1115-1166. https://doi.org/10.1177/0149206311432640

Piraveenan, M., Prokopenko, M., & Hossain, L. (2013). Percolation centrality: Quantifying graph-theoretic impact of nodes during percolation in networks. *PLoS One*, 8(1), e53095. https://doi.org/10.1371/journal.pone.0053095

Polanyi, M. (2009). *The tacit dimension*. University of Chicago Press.

Popescul, D. (2012, June 9-10). Knowledge flows percolation model – a new model for the relation between knowledge and innovation. In A. Mohammad, M. Syazli, & R. Shakila (Eds.), *Proceedings of the 18th International Business Information Management Association Conference (Innovation and Sustainable Economic Competitive Advantage: From Regional Development to World Economies)* (pp. 445-453). International Business Information Management Association.

Roque, L. (2004). *Contribuição para uma engenharia do contexto* [Contribution to context engineering] [PhD thesis, Universidade de Coimbra, Coimbra, Portugal]. https://estudogeral.sib.uc.pt/handle/10316/1743

Roque, L., Almeida, A., & Figueiredo, A. D. (2004, June 14-16). Context engineering: An IS development research agenda. *Proceedings of the 12th European Conference on Information Systems* (pp. 1582-1595). Turku School of Economics and Business Administration.

Sahimi, M. (1994). *Applications of percolation theory.* Taylor & Francis.

Schmidt, B. (2006). Proof of principle studies. *Epilepsy Research*, 68(1), 48-52. https://doi.org/10.1016/j.eplepsyres.2005.09.019

Schmidt, L. K. (2012). *Hermenêutica* [Hermeneutics]. Editora Vozes.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.

Stauffer, D., & Aharony, A. (1992). *Introduction to percolation theory*. (2nd ed.) Taylor & Francis. https://doi.org/10.1201/9781315274386

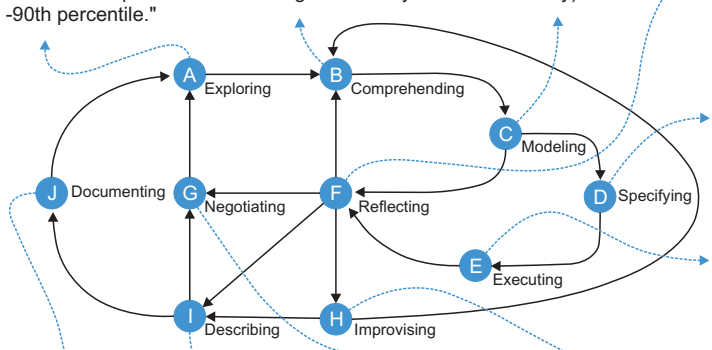
APPENDIX 1. Some examples of PoC scenarios

"All efficiency enabled, 30 KB block size, 10% reads, 90% writes, 145K IOPS, 94% random IO, <0.88 ms response time -90th percentile."

"Where do you want the 1 ms [latency] to be measured? At the application layer or storage device layer?"

"But for each experiment, we built a wrapper that will analyze and average all the results (IOPS and latency)."

"Are they really going to use that configuration in real life? The contradictory information we found is that all their scripts have 1,024 outstanding I/O."



"They would like to use its internal DB scripts to exercise the IO. Why do we also need to run other tools with an insane block size that does not represent that DB?"

"After that, I started the same tests. We just got 142 K with 1.46 ms read latency. They suspect there is a problem with the quipment."

"With Vergonha-IO [anonymized name], we were able to see 32 GB/s reads. We tried to model using the same parameters used by your application and we cannot pass 22 GB/s reads."

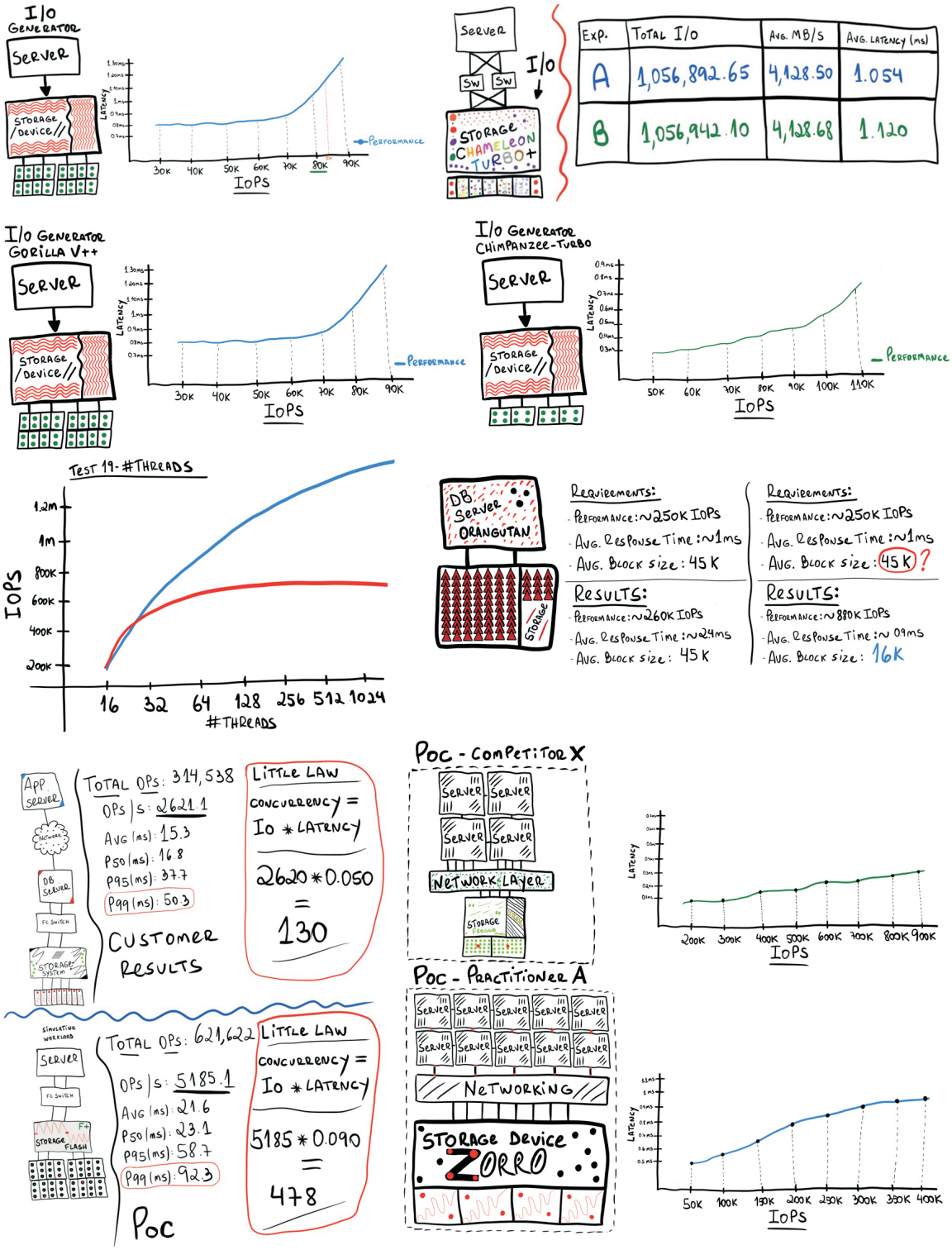
"Right now, customer-based tests with this specific version of the workload tool are substantially different compared with the tool's predecessor. Please remember that IOPS are not all created in an equal manner."

"If we present these finds to Mr. Goofy [anonymized name] and show him that our platform is topping out at similar speeds using different tools. Maybe that will be enough for them to move the needle."

"Doing some kernel instrumentation, we identified this software does not open all files with O_DIRECT flag. Thus, all the IO are staying on the filesystem cache, not moving the bottleneck to the storage engine. Do you know if the goal is to test reads or writes or both?"

 https://doi.org/10.1633/JISTaP.2021.9.1.1

APPENDIX 2. Some examples of PoC sketches



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

RESEARCH PAPER
J Inf Sci Theory Pract 9(1): 24-34, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.1.2

Identification of Profane Words in Cyberbullying Incidents within Social Networks

Wan Noor Hamiza Wan Ali
Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: p93244@siswa.ukm.edu.my

Masnizah Mohd*
Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: masnizah.mohd@ukm.edu.my

Fariza Fauzi
Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: fariza.fauzi@ukm.edu.my

ABSTRACT

The popularity of social networking sites (SNS) has facilitated communication between users. The usage of SNS helps users in their daily life in various ways such as sharing of opinions, keeping in touch with old friends, making new friends, and getting information. However, some users misuse SNS to belittle or hurt others using profanities, which is typical in cyberbullying incidents. Thus, in this study, we aim to identify profane words from the ASKfm corpus to analyze the profane word distribution across four different roles involved in cyberbullying based on lexicon dictionary. These four roles are: harasser, victim, bystander that assists the bully, and bystander that defends the victim. Evaluation in this study focused on occurrences of the profane word for each role from the corpus. The top 10 common words used in the corpus are also identified and represented in a graph. Results from the analysis show that these four roles used profane words in their conversation with different weightage and distribution, even though the profane words used are mostly similar. The harasser is the first ranked that used profane words in the conversation compared to other roles. The results can be further explored and considered as a potential feature in a cyberbullying detection model using a machine learning approach. Results in this work will contribute to formulate the suitable representation. It is also useful in modeling a cyberbullying detection model based on the identification of profane word distribution across different cyberbullying roles in social networks for future works.

Keywords: cyberbullying, profane words, cybercrime, harassment, social network, machine learning

Received: October 19, 2020 **Accepted:** February 21, 2021

***Corresponding Author:** Masnizah Mohd
https://orcid.org/0000-0001-8908-8755
E-mail: masnizah.mohd@ukm.edu.my

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far 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.

© Wan Noor Hamiza Wan Ali, Masnizah Mohd, Fariza Fauzi, 2021

1. INTRODUCTION

The rapid development of information and communication technologies and the great proliferation of social networking sites (SNS) have made social media a popular socialization tool, especially among teenagers. Since the introduction of these SNS such as Facebook, Instagram, Twitter, and YouTube, they have been able to attract the attention of millions of users to the point where these SNS have become a necessity in life (Tarmizi et al., 2020). According to Boyd and Ellison (2007), social networks are defined as web-based services that allow individuals to build public or semi-public profiles in a limited system, and to clearly know the list of other users with whom they connect as well as to view their contact lists. Thus, the use of the Internet as a medium of communication makes SNS an important socialization tools to share information, exchange opinions, interact with each other, and to conduct timeless and borderless online business (Zainudin et al., 2016). However, SNS could give negative implications if people use SNS for cybercrime purposes (Sheeba et al., 2019).

1.1. Cyberbullying Detection

One of the cybercrimes that usually occur on SNS is cyberbullying (Nandhini & Sheeba, 2015). Cyberbullying has been defined in multiple ways but have the same characterization where it takes into account the people involved, the kind of medium, and how cyberbullying occurs. Hinduja and Patchin (2018) defined cyberbullying as "willful and repeated harm inflicted through the medium of electronic text." Based on this definition, they did not mention cyberbullying that happened based on the specific age group of users, in contrast to other researchers who believe that cyberbullying happens only among students such as children and teenagers (Sutton, 2011). When it comes to adults, cyberbullying can be defined as cyber harassment or cyberstalking, besides cyber incivility (Giumetti et al., 2012; Sutton, 2011). However, cyberbullying is an aggressive behavior by an individual or group of people known as harassers towards another individual (i.e., victim) using the Internet as a medium. Harassers are able to use various ways to bully the victim such as spreading rumors and spreading information related to the victim for the purpose of embarrassing him/her. The people involved in cyberbullying incidents can be classified as a harasser, victim, assistant, defender, reinforcer, and accuser (Salmivalli, 1999; Salmivalli et al., 1996; Xu et al., 2012). Table 1 shows the description of these six roles. These individuals may play different roles in a single cyberbullying incident (Xu et al., 2012). Van Hee et al. (2018) only consider four roles: 1) harasser, 2) victim, 3) assistant, and 4) defender. They refer to the assistant and defender in cyberbullying incidents as bystanders (specifically, bystander-assistant and bystander-defender, respectively).

The impact and frequency of cyberbullying incidents can intensify or lessen depending on the roles involved (Moxey & Bussey, 2020). The existence of the roles beside harasser and victim has been recognized as important (Bastiaensens et al., 2014; Salmivalli, 2010). If a cyberbullying incident involves only the harasser and victim, the issue is likely to be cut short. However, if the incident involves roles such as assistant and reinforcer, this situation can worsen and become more dangerous, as the harasser has received support from other online social network users, and does not consider the cyberbullying wrong, and will continue to threaten the victim. Nonetheless, if there are users playing the role of defender involved, it is likely that the incident will not be extended because the harasser has no support from the other online social network users (Salmivalli, 2010). Thus, individual involvement in cyberbullying incidents is noteworthy in cyberbullying detection.

Discriminative features are required to develop a cyberbullying detection model by using a supervised machine learning approach. In general, basic steps in modeling the

Table 1. Description of roles involved in cyberbullying

Role	Description
Harasser	Person who did the bullying across online social networks
Victim	Person who was the target of harasser
Assistant	Helps harasser to do the bullying
Defender	Defends victim from harasser
Reinforcer	Indirectly involved in cyberbullying but encourages harasser and gives impetus for continuation, such as cheering or laughing
Accuser	Accusing someone as the harasser

cyberbullying detection are started from data collection, data annotation, data pre-processing, feature extraction, and classification (Talpur & O'Sullivan, 2020). However, collecting data from online social networks is not an easy task. According to Salawu et al. (2017), the challenge lies in the form of privacy and ethical matters which most researchers have to face before the data can be harvested, especially when it involves data related to user information such as age and gender. Besides this, different online social networks provide different metadata, and researchers should determine the detection task before selecting the data source to be used. However, the corpus should be annotated after collecting the same from social networks. The data annotation is the process where raw data will be identified into meaningful and informative annotation to classify text. Meanwhile, data pre-processing is one of the crucial steps in cyberbullying detection because we need to remove outliers and standardize the data before it can be used to build the cyberbullying model during classification process. A few examples of data pre-processing are tokenization, stemming, lemmatization, remove numbering, and stopwords removal.

Feature extraction in cyberbullying detection is dependent on the corpus data that have been harvested. As an example, the corpus may or may not contain metadata of elapsed time between comments, followers, following, time of posting, and other data. These metadata may be required to identify the interaction between users of the online social network site. Salawu et al. (2017) categorized four types of features used in cyberbullying detection: 1) content-based (e.g., cyberbullying keywords, profanity, pronouns, and n-grams), 2) sentiment-based (e.g., emoticons), 3) network-based (e.g., number of following, number of followers), and 4) user-based (e.g., age, gender). Content-based features are commonly used features by researchers in cyberbullying detection. In addition, the profanity feature is the highest feature that has been used across cyberbullying detection studies as compared to other features in the content-based category, since cyberbullying is associated to an aggressive behavior (Salawu et al., 2017). Finally, we use an algorithm as classifier in the classification process as the final step for cyberbullying detection. Common algorithms usually implemented in supervised machine learning are the Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, and Linear Regression (Muneer & Fati Mohamed, 2020).

However, state-of-the-art of cyberbullying detection using a machine learning approach has focused on profane words as a feature rather than utilizing the profane features according to four roles (harasser, victim, assistant, and defender) involved in cyberbullying incidents. This has motivated us to identify profane words from the corpus to analyze the word distribution based on four roles involved in cyberbullying incidents within social network. However, a limitation of this work is that we only identify lexical profane words in accordance to a profane word dictionary. We addressed the following research questions:

1. Which role used profane words with the highest frequency based on word distribution?
2. What are the commonly used words by users in cyberbullying incidents?

We hypothesize that identifying word distribution for each role involved in cyberbullying detection can give useful insights for us to better extract profanity features from the corpus for use in the classification process. Results from the experiment will guide us for future work in cyberbullying detection using a machine learning approach. The contribution of this work is noteworthy, as it focuses on profane words distribution based on four different roles involved in cyberbullying detection. The results from this work will be implemented in formulating the representation of features in order to develop a model of cyberbullying detection. We also combined two different external sources of profane word dictionaries with the purpose to strengthen the results of the experiment.

The remainder of this paper is structured as follows: Section 2 presents the related work on automatic cyberbullying detection using profanity as a feature. The methodology in Section 3 describes the workflow for the identification of profane words. Next, we present the results from the experiment in Section 4. Meanwhile, Section 5 discusses the analysis of profane word distribution. Finally, Section 6's conclusion and future work concludes this paper.

2. RELATED WORK

Nowadays, profane words are commonly used in daily face to face conversations as well as on the Internet, especially in online social networks (Laboreiro & Oliveira, 2014). A profane word is a curse word that is used with offensive meaning and usually used in cyberbullying incidents and hate speech (Laboreiro & Oliveira, 2014). Profanity is a common feature used in cyberbullying detection. State-of-the-art works on cyberbullying detection that used profanity feature to detect cyberbullying in on-

line social networks include research works by Dinakar et al. (2011), Van Hee et al. (2018), Kontostathis et al. (2013), and Al-garadi et al. (2016).

Dinakar et al. (2011) have focused on cyberbullying detection based on textual communication that involved three main sensitive topics, which are sexual, race/culture, and intelligence. By using a corpus from YouTube, they annotated the corpus that consists of negative and profane words according to those three topics. However, they stated that if a comment in a YouTube video contains negative or profane words, it does not necessarily mean that the comment can be labeled as a cyberbullying incident. For example, "I'm disgusted by what you said today, and I never want to see you again" is difficult to classify as cyberbullying even though that sentence indicates the lexicon 'disgusted.'

Profane words are also identified by Al-garadi et al. (2016) in cyberbullying detection because they believe it is a sign of offensive behavior. They extracted profane words based on a dictionary of profanities which is compiled in previous research works by Reynolds et al. (2011) and Wang et al. (2014). In addition, a study by Chatzakou et al. (2017) uses an external resource, the 'hatebase' database (https://hatebase.org) to determine the 'hateful' score for the collected tweets on a scale [0,100], where 0 indicates no hatred and 100 indicates hatred. However, they stated that profane words are not suitable for classifying tweets as hateful or aggressive cyberbullying, as the tweets are short and commonly comprise altered words, emoticons, and URLs. So, they concluded that users who like to harass other users have a tendency to use profane words, but they are not much different from normal user behavior.

The Google profanity list (https://code.google.com/archive/p/badwordslist/downloads) is used as a lexicon dictionary for feature extraction by Van Hee et al. (2018). However, Van Hee et al. (2018) stated that profanity has a low percentage in F1 measure but it still can be a useful feature if it is integrated with other features such as sentiment, n-grams (character), and topic model.

Tarmizi et al. (2020) used the corpus from Twitter to detect cyberbullying activities. The profane words extracted from the 50 Twitter users' corpus is used to track the activities by the users. The extraction is based on an external source, www.noswearing.com. In the classification process, about 10,183 from 29,701 profane words occurrences are misclassified. It may be due to the low number of tweet activities having a high number of profane words and the high number of tweet activities having a low number of profane words.

Current findings showed that profane words were highly used as a feature to detect the cyberbullying that happened in social networks. Nonetheless, there is no research that considers frequency of profane words based on roles involved in cyberbullying incidents. This is important to ensure that feature use in classification processes can build a robust model for cyberbullying detection.

3. METHODOLOGY

We set up an experiment to test the aforementioned hypothesis and answer the research questions. Thus, in this section a workflow to analyze profane word distribution across the different cyberbullying roles in the corpus is proposed as illustrated in Fig. 1. The workflow involves four steps: 1) dataset, 2) data preparation, 3) data preprocessing, and 4) profane word extraction.

3.1. Dataset

The corpus used in this work is the AMiCA Bullying Cyber Dataset, a secondary dataset obtained from Van Hee et al. (2018). The corpus is collected from ASKfm's social network site from April to October 2013. It consists of English and Dutch language. However, only the English language is used in this study. ASKfm is one of the most popular online social networking platforms for teens and adults where users can interact without having to reveal their own identities (Ashktorab et al., 2017). The corpus obtained is in the form of annotations annotated using Brat Rapid Tool Annotation (BRAT) (Stenetorp et al., 2012).

In the annotation process as described in Van Hee et al. (2018), messages collected are presented to the annotators in chronological order within the context of the original content of the conversation. One conversation consists of a series of messages from two or more users. In each conversation, each message is preceded by a ¶ token. The token is intended to facilitate and speed up the annotation process. Firstly, the nature of each message is taken into account whether it is harmful or otherwise. The level of harmfulness is indicative of cyberbullying. Hence, for each message, the annotators determine whether the message contains indication of cyberbullying or not based on three types of scales, namely:

1. Harmfulness Score=0: Messages that do not contain cyberbullying or indication of cyberbullying.
2. Harmfulness Score=1: Message containing cyberbullying or an indication of cyberbullying.

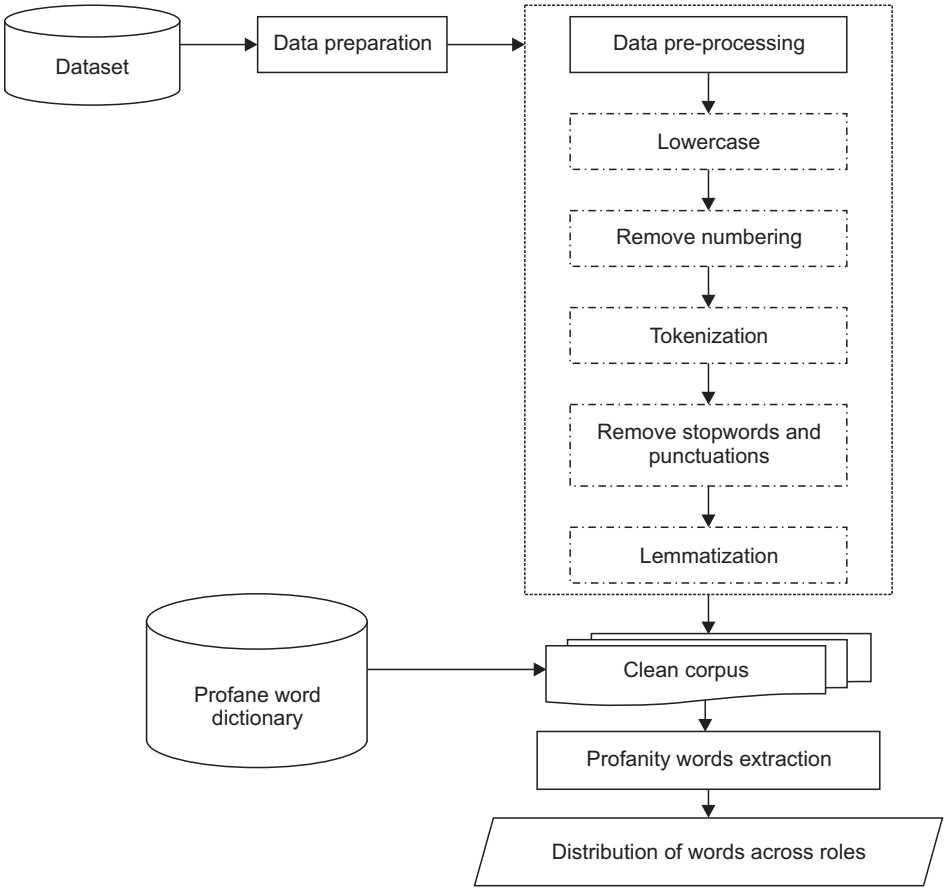


Fig. 1. Workflow of profane words extraction.

3. Harmfulness Score=2: Messages that contain a serious indication of cyberbullying (e.g., death threats).

Secondly, if the message is rated with a harmfulness score of 1 or 2, the annotators must specify the user's role for the messages in the following format: *score_role* i.e., *1_harasser, 2_harasser, 1_victim, 2_victim, 1_bystander_assistant, 2_bystander_assistant, 1_bystander_defender,* or *2_bystander_defender.*

In this work, profane words are extracted from four different roles without any consideration for the harmfulness score. For example, both *1_harasser* and *2_harasser* are simply considered as *harasser* for the extraction of profane words.

3.2. Data Preparation

Data preparation processing is performed upon obtaining the dataset. All files containing the original messages and annotations are unified in one folder. The BRAT reader program is imported and implemented in the Anaconda Spyder platform using Python 3.7 programming language, and the original Python codes from https://github.com/clips/bratreader/tree/master/bratreader are adapted to parse the original messages and annotations. BRAT reader is a Python program for reading and understanding BRAT repositories and for storing files in XML format for easy access (Computational Linguistics Research Group, 2019). The program was developed by Computational Linguistics & Psycholinguistics from the University of Antwerp, Belgium. The original messages and annotation files are loaded into .xml files before being stored and converted to .csv format using the BRAT reader program. Then we imported the converted files into the Anaconda Spyder platform for data pre-processing.

The corpus is then processed for better understanding. Table 2 shows the descriptive statistics of the corpus. The corpus has a total of 120,512 posts. They were classified into two types of post which are cyberbullying posts (5,382) and non-cyberbullying posts (115,130). For the cyberbullying posts, the number of posts is calculated for each of the four roles defined in a cyberbullying incident. As can be observed from Table 2, the corpus is imbalanced (only 4.47% of the corpus are labelled as cyberbullying posts) and this situation will affect the performance

 https://doi.org/10.1633/JISTaP.2021.9.1.2

of standard classifiers and predictive models. This issue will be addressed in our future work.

3.3. Data Pre-Processing

Based on the corpus provided, each original message goes through numbers removal and lowercase conversion. Then, all punctuation and stopwords are removed from the corpus. The next step is the tokenization of the corpus by converting each sentence in the message into a set of words known as tokens for easier comprehension (Sekharan et al., 2018). Finally, lemmatization is performed to obtain the root word for each token. Table 3 shows the pre-processing steps of the original data and processed output using the following example sentence: "NIGGA you look like you are a 60 years old hag, even SLUTS like you!!!"

3.4. Profane Word Extraction

We utilized two different external sources to identify profane words from the corpus according to the four different roles involved in cyberbullying, as follows:

1. Curse word lexicon from https://www.cs.cmu.edu/~biglou/resources/bad-words.txt, that consists of 1,383 lexicons (Zois et al., 2018)
2. Google profanity list https://code.google.com/archive/p/badwordslist/downloads, which consists of 458 lexicons (Van Hee et al., 2018)

Combining both external sources gives 1,841 lexicons and we compiled this into one file before identifying the profane words. The identification of profane words from the corpus is conducted in a controlled environment on a single machine. The reason is to make sure of consistency of performance in order to produce results. We performed countvectorizer for feature extraction, which is used to convert a corpus to a matrix of tokens by making use of Scikit-learn (Pedregosa et al., 2011). Then, we implemented fit_transform of lexicon dictionary, and word distribution for each role is calculated.

4. RESULTS

In this section, we present the findings from an experiment of profane word identification as a part of feature extraction, which could be used in the classification process of cyberbullying detection. Table 4 shows the percentage of profane words distribution across four roles in cyber-

Table 2. Descriptive statistics of ASKfm corpus

Post	Corpus size	Percentage of corpus size (%)	Score_Role	Number of post	Total post each role
Non-cyberbullying post	115,130	95.53	-	-	-
Cyberbullying post	5,382	4.47	1_Harasser	2,886	2,886+693=3,579
			2_Harasser	693	
			1_Victim	1,276	1,276+79=1,355
			2_Victim	79	
			1_Bystander_Defender	391	391+33=424
			2_bystander_Defender	33	
			1_Bystander_Assistant	17	17+7=24
			2_Bystander_Assistant	7	

Table 3. Output for each pre-processing step

Pre-processing steps	Output for pre-processing
Step 1: Remove numbering	NIGGA you look like you are a 60 years old hag, even SLUTS like you!!!
Step 2: Lowercase	nigga you look like you are a years old hag, even sluts like you!!!
Step 3: Remove punctuations and stopwords	nigga hag sluts
Step 4: Tokenization	['nigga', 'hag', 'sluts']
Step 5: Lemmatization	['nigga', 'hag', 'slut']

bullying incidents. A formula to calculate the percentage of profane word distribution is built. The number of profane words indicated the total number of profane words that have been extracted from tokens for each role. The formula as follows:

$$\text{Percentage Profanity Words Distribution} = \frac{\text{Number of Profane Words}}{\text{Total Number of Tokens}} \times 100\%$$

The Harasser role consists of 12,958 tokens from 3,579 posts and total profane words used are 3,830. This yields a percentage of word distribution of about 29.56%. This shows that profane words used by harassers are more than one quarter of the overall tokens. For the victim role, 152 profane words are used over 6,261 tokens contained in 1,355 posts, which yields a percentage of word distribution of about 2.43%. On the other hand, the percentage for the bystander_defender role is about 19.88%. This is calculated from a 730 number of profane words from 3,672 numbers of tokens contained in 424 posts. For bystander_assistant, the usage of profane words is 21 from an 88 number of tokens contained in 24 posts. This makes the percentage of profane word distribution for this role about 23.86%.

Fig. 2 presents findings of the top 10 most common profane words in four different roles in cyberbullying incidents. The graph incorporates the information about total numbers of profane words used by roles (x-axis). In the y-axis is shown the top 10 profane words used by each role in cyberbullying incidents. The profane word 'fuck' is the highest word used in all four roles in cyberbullying incidents. The 'fuck' word used by harassers is about 720, 368 is used by victims, 200 by bystander_defender, and 5 by bystander_assistant, thus making the harasser ranked first compared to others.

For harasser, profane words are ranked as follows: fuck (720 profane words), bitch (274 profane words), ugly (216 profane words), hate (173 profane words), shit (131 profane words), ass (113 profane words), cunt (111 profane words), dick (111 profane words), slut (111 profane words), and faggot (104 profane words). Victim profane words in rank are fuck (368 profane words), shit (92 profane words), bitch (88 profane words), cunt (71 profane words), dick (40 profane words), stupid (34 profane words), ass (31 profane words), pussy (28 profane words), fat (27 profane words), and dumb (22 profane words). For bystander_defender, profane words used in rank are fuck (200 profane words), shit (59 profane words), bitch (44 profane words), cunt (43 profane words), fat (32 profane words), ass (22 profane words), slut (21 profane words), stupid (18 profane words), stfu (18 profane words), and twat (16 profane words). Hence, bystander_assistants also used profane words in their cyberbullying incidents even though the total numbers for each word are less compared to other roles. The rank for bystander_assistant profane words are fuck (5 profane words), cunt (3 profane words), slut (2 profane words), lmfao (2 profane words), piss (1 profane word), bitch (1 profane word), fag (1 profane word), boob (1 profane word), damn (1 profane word), and hole (1 profane word).

We present Fig. 3 by summarizing Fig 2, which shows five highly used profane words in four roles in cyberbullying incidents. The x-axis contains information of top five profane words that are commonly used in four roles, while the y-axis displays the number of each profane words used by four roles. The total number of the 'fuck' word used by four roles is 1,293, which is the highest compared to other profane words. Secondly, the 'bitch' word used in cyberbullying incidents is 407, followed by cunt (228), slut (156), and twat (82).

5. DISCUSSION

Our focus in this work is to identify profane words used by the aforementioned roles in cyberbullying incidents using a lexicon profane word dictionary in order to analyze word distribution. This is to allow extraction of the profanity features to be implemented into the classification process of a cyberbullying detection model, and at the same time to answer the research questions. Findings revealed that each role used profane words in their con-

Table 4. Percentage of profane word distribution across four roles in cyberbullying incidents

Roles	Number of profane words	Total number of tokens	Percentage profane words distribution (%)
Harasser	3,830	12,958	29.56
Victim	152	6,261	2.43
Bystander_Defender	730	3,672	19.88
Bystander_Assistant	21	88	23.86

 https://doi.org/10.1633/JISTaP.2021.9.1.2

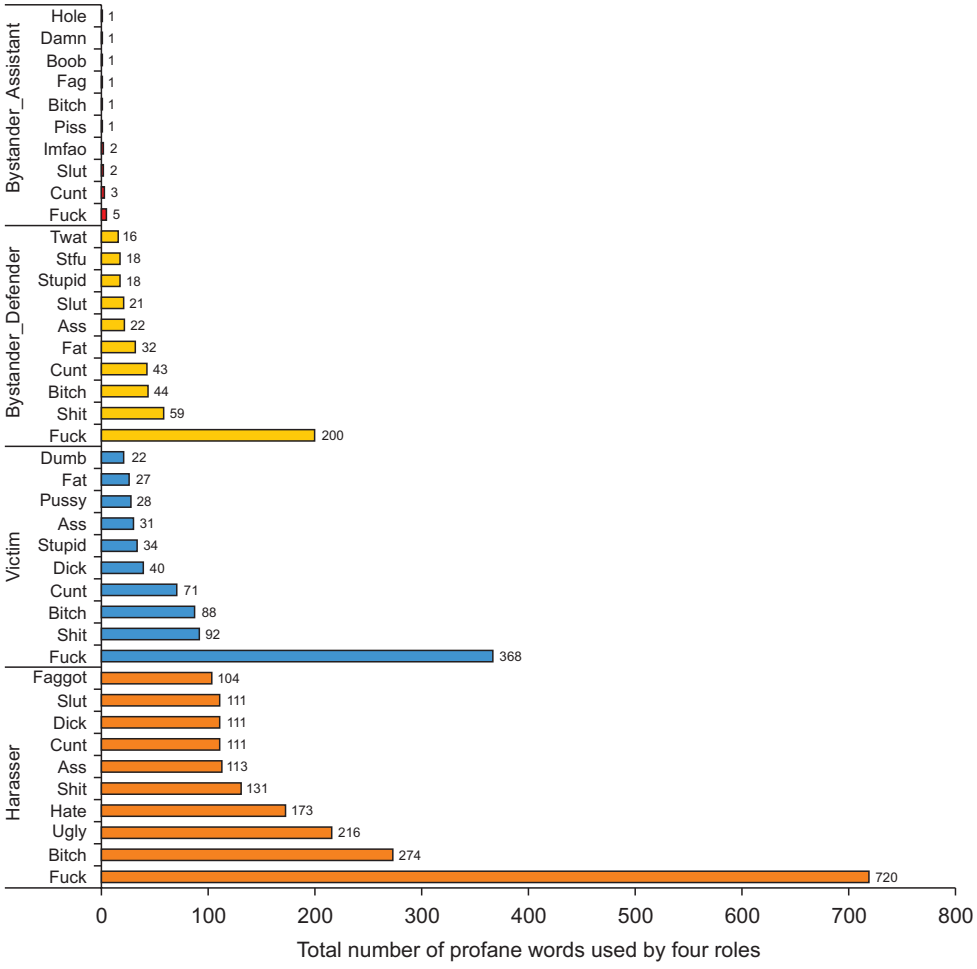
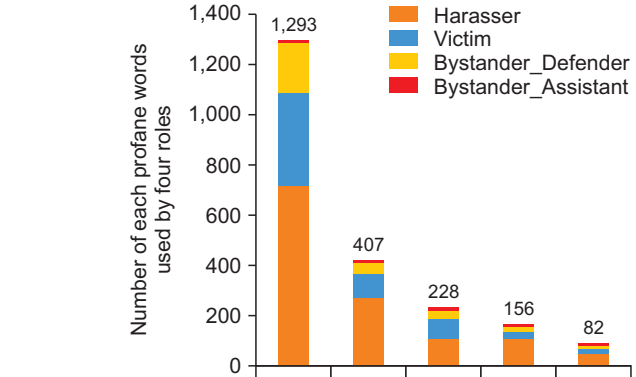


Fig. 2. Top 10 profane words distribution.



	Fuck	Bitch	Cunt	Slut	Twat
Bystander_Assistant	5	1	3	2	1
Bystander_Defender	200	44	43	21	16
Victim	368	88	71	22	17
Harasser	720	274	111	111	48

Fig. 3. Top 5 most common profane words.

version which were declared as containing a cyberbullying indication. This shows profane words can be the profanity feature in cyberbullying classification.

As mentioned, a harasser is the person who does the bullying towards victims, and profane words are always used by them to hurt others' feelings. This is to justify why the harasser is the identified role which used profane words higher than any other roles did. Meanwhile, victims may have two different reactive behaviors; either they become aggressive or become passive when being confronted by a harasser in a cyberbullying incident. Aggressive behavior appears when they try to fight back against the harasser by using the same kind of language as in the profane words, for example, "Shut the fuck up you smelly." On the other hand, a victim can react passively where he/she accepts the insults from the harasser or just replies without using any profane words, e.g., "Stop, just leave me alone." The usage of profane words by the victim could be attributed to the victim's response where the victim may

behave aggressively when confronted by the harasser; for example, "Leave Amber alone, don't be jealous and accept the fact that Amber has other friends, even if they are hipster;) so shut the fuck up u lot will never be the new Megan sweeney so positive vibes motherfuckers." Because of that, the victim is a role that also used profane words in conversations even though the percentage yielded is very low. We can see that bystander_defenders used the terms 'fuck' and 'motherfucker,' but in a positive way where they try to stop the harasser from bullying the victim, and this conforms to its definition whereby a bystander_assistant supports the harasser, likely with similar aggressive behavior to the harasser (Van Hee et al., 2018). Furthermore, it is observed that the posts by the bystander_assistant are similar to that of the harasser. This issue makes cyberbullying detection in online social networks a difficult task, especially in the role determination stage. As a whole, we can declare that the first research question is answered.

As discussed in Teh et al. (2018), the term 'fuck' has the highest score compared to other terms in year 2017. Thus, we can conclude that this kind of profane word has long been used before 2017, as this corpus was collected in 2013. Besides this, the terms 'bitch' and 'cunt' are also in the top 10 most common words used in every role in cyberbullying incidents. However, the total numbers of each profane word for each role are much different. This is caused by an imbalance of data under cyberbullying posts as mentioned in the sub-section "Data Preparation." Cyberbullying posts annotated as harrasser, victim, bystander_assistant, and bystander_defender roles have different number gaps as described in Table 2. Besides this, all of these terms are categorized as profane words and could be one of the features in a cyberbullying detection model, yet their use does not necessarily indicate the occurrence of cyberbullying; as an example, "I'm tired as fuck and have to work tomorrow." This is concluded to mean that the second research question is answered. Then, there are words used by users by substituting the original letter with another letter. For example, the letter 'u' is substituted by letter 'v' in order to convert the word 'fuck' to 'fvck,' and letter 'b' is substituted by 'v' to convert a word from 'bitch' to 'vitch.' Other than that, there are also misspelled words such as 'fck' where the original word is 'fuck.' When using a profanity dictionary to identify profane words, misspelled words cannot be identified and are considered as non-profane words. This indicates that there is a probability for a classifier to mistakenly detect cyberbullying if it only depends on profanity as a feature. However, with the information provided by Fig. 3, we can consider word embedding to improve the efficiency of the classifier for cyberbullying detection.

6. CONCLUSION AND FUTURE WORK

Recently, cyberbullying has been acknowledged as a serious national health issue in the midst of online social network users. One of the ways to solve this problem is by developing a model to detect cyberbullying in social media. Thus, this paper focuses on extracting profane words from a state-of-the-art corpus with four different roles in order to analyze profane word distribution. From the results, all of the four roles used profane words in their conversation and we concluded that users tend to use them when they are involved in cyberbullying. However, the use of profane words does not necessarily indicate cyberbullying incidents. This issue could be overcome by looking at the underlying weightage and distribution; thus, word embedding for text representation can be considered. Also, we can conclude that profane words can be one of the features used in order to develop a cyberbullying model but with the presence of other features to further improve the accuracy of the model.

In our future work, we will extract and incorporate other features to be integrated with the profanity feature in the classification process. Other than that, oversampling technique can be implemented in this corpus for a balance dataset before undergoing the classification process. Imbalance data may affect the classifier performance in training data to detect cyberbullying.

Lexical variation will also be explored to increase efficiency in the classification process of cyberbullying detection. The distribution of profane words in this result can be used in the future for an early cyberbullying detection model.

CONFLICTS OF INTEREST

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

REFERENCES

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 63, 433-443. https://doi.org/10.1016/j.chb.2016.05.051

Ashktorab, Z., Golbeck, J., Haber, E., & Vitak, J. (2017, June

25-28). Beyond cyberbullying: Self-disclosure, harm and social support on ASKfm. In P. Fox, D. McGuinness, & L. Poirer (Eds.), *WebSci '17: Proceedings of the 2017 ACM Web Science Conference* (pp. 3-12). ACM. https://doi.org/10.1145/3091478.3091499

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior,* 31, 259-271. https://doi.org/10.1016/j.chb.2013.10.036

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210-230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017, June 25-28). Mean birds: Detecting aggression and bullying on Twitter. In P. Fox, D. McGuinness, & L. Poirer (Eds.), *WebSci '17: Proceedings of the 2017 ACM Web Science Conference* (pp. 13-22). ACM. https://doi.org/10.1145/3091478.3091487

Computational Linguistics Research Group. (2019). *Bratreader*. https://github.com/clips/bratreader

Dinakar, K., Reichart, R., & Lieberman, H. (2011, July 17-21). Modeling the detection of textual cyberbullying. In N. Nicolov & J. G. Shanahan (Eds.), *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 11-17). AAAI Press.

Giumetti, G. W., McKibben, E. S., Hatfield, A. L., Schroeder, A. N., & Kowalski, R. M. (2012). Cyber incivility @ work: The new age of interpersonal deviance. *Cyberpsychology, Behavior and Social Networking*, 15(3), 148-154. https://doi.org/10.1089/cyber.2011.0336

Hinduja, S., & Patchin, J. W. (2018). *Cyberbullying: Identification, prevention, and response*. https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2018.pdf.

Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L. (2013, May 2-4). Detecting cyberbullying: Query terms and techniques. In H. Davis, H. Halpin, & A. Pentland (Eds.), *WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference* (pp. 195-204). ACM. https://doi.org/10.1145/2464464.2464499

Laboreiro, G., & Oliveira, E. (2014, October 6-8). What we can learn from looking at profanity. In A. Baptista, N. Mamede, S. Candeias, I. Paraboni, T. A. S. Pardo, & M. G. V. Nunes (Eds.), *Computational Processing of the Portuguese Language: 11th International Conference, PROPOR 2014. Proceedings* (pp. 108-113). Springer. https://doi.org/10.1007/978-3-319-09761-9

Moxey, N., & Bussey, K. (2020). Styles of bystander intervention in cyberbullying incidents. *International Journal of Bullying Prevention*, 2(1), 6-15. https://doi.org/10.1007/s42380-019-00039-1

Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. *Future Internet*, 12(11), 187. https://doi.org/10.3390/fi12110187

Nandhini, B. S., & Sheeba, J. I. (2015, March 6-7). Cyberbullying detection and classification using information retrieval algorithm. In S. A. Khadar (Ed.), *ICARCSET '15: Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)* (pp. 1-5). ACM. https://doi.org/10.1145/2743065.2743085

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. The *Journal of Machine Learning Research*, 12, 2825-2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December 18-21). Using machine learning to detect cyberbullying. In X. Chen, T. Dillon, H. Ishbuchi, J. Pei, H. Wang, & M. A. Wani (Eds.), *Proceedings of the 10th International Conference on Machine Learning and Applications (ICMLA 2011)* (pp. 241-244). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICMLA.2011.152

Salawu, S., He, Y., & Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1), 3-24. https://doi.org/10.1109/TAFFC.2017.2761757

Salmivalli, C. (1999). Participant role approach to school bullying: Implications for interventions. *Journal of Adolescence*, 22(4), 453-459. https://doi.org/10.1006/jado.1999.0239

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior,* 15(2), 112-120. https://doi.org/10.1016/j.avb.2009.08.007

Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior*, 22(1), 1-15. https://doi.org/10.1002/(SICI)1098-2337(1996)22:1%3C1::AID-AB1%3E3.0.CO;2-T

Sekharan, S. C., Vadivu, G., & Rao, M. V. (2018). A comprehensive study on sarcasm detection techniques in sentiment analysis. *International Journal of Pure and Applied Math-*

ematics, 118(22), 433-442. https://acadpubl.eu/hub/2018-118-22/articles/22a/63.pdf

Sheeba, J. I., Devaneyan, S. P., & Cadiravane, R. (2019). Identification and classification of cyberbully incidents using bystander intervention model. *International Journal of Recent Technology and Engineering (IJRTE),* 8(2S4), 1-6. https://doi.org/10.35940/ijrte.B1001.0782S419

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012, April 23-27). BRAT: A web-based tool for NLP-assisted text annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102-107). ACL. https://dl.acm.org/doi/10.5555/2380921.2380942

Sutton, B. B. (2011). Cyberbullying: *An interview with Parry Aftab*. https://etcjournal.com/2011/02/17/7299

Talpur, B. A., & O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. *PLoS One*, 15(10), e0240924. https://doi.org/10.1371/journal.pone.0240924

Tarmizi, N., Saee, S., & Ibrahim, D. H. A. (2020). Detecting the usage of vulgar words in cyberbully activities from Twitter. *International Journal on Advanced Science, Engineering and Information Technology*, 10(3), 1117-1122. http://doi.org/10.18517/ijaseit.10.3.10645

Teh, P. L., Cheng, C., & Chee, W. (2018, March 23-25). Identifying and categorising profane words in hate speech. In A. Gokhale & S. Zhang (Eds.), *Proceedings of 2018 the 2nd International Conference on Compute and Data Analysis* (pp. 65-69). ACM. https://doi.org/10.1145/3193077.3193078

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLoS One*, 13(10), e0203794. https://doi.org/10.1371/journal.pone.0203794

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014, February 15-19). Cursing in English on twitter. In S. Fussell & W.e Lutters (Eds.), *CSCW'14: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 415-425). ACM. https://doi.org/10.1145/2531602.2531734

Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012, June 3-8). Learning from bullying traces in social media. In J. Chu-Carroll (Ed.), *NAACL HLT '12: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 656-666). ACL. https://dl.acm.org/doi/10.5555/2382029.2382139

Zainudin, N. M., Zainal, K. H., Hasbullah, N. A., Wahab, N. A., & Ramli, S. (2016, May 16-17). A review on cyberbullying in Malaysia from digital forensic perspective. *ICICTM '16: Proceedings of 1st International Conference on Information and Communication Technology* (pp. 246-250). IEEE. https://doi.org/10.1109/ICICTM.2016.7890808

Zois, D., Kapodistria, A., Yao, M., & Chelmis, C. (2018, April 15-20). Optimal online cyberbullying detection. In M. Hayes & H. Ko (Eds.), *Proceedings of 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 2017-2021). IEEE. https://doi.org/10.1109/ICASSP.2018.8462092

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

RESEARCH PAPER
J Inf Sci Theory Pract 9(1): 35-53, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.1.3

Topic Modeling and Sentiment Analysis of Twitter Discussions on COVID-19 from Spatial and Temporal Perspectives

Iyad AlAgha*
Faculty of Information Technology, The Islamic University of Gaza, Gaza, Palestine
E-mail: ialagha@iugaza.edu.ps

ABSTRACT

The study reported in this paper aimed to evaluate the topics and opinions of COVID-19 discussion found on Twitter. It performed topic modeling and sentiment analysis of tweets posted during the COVID-19 outbreak, and compared these results over space and time. In addition, by covering a more recent and a longer period of the pandemic timeline, several patterns not previously reported in the literature were revealed. Author-pooled Latent Dirichlet Allocation (LDA) was used to generate twenty topics that discuss different aspects related to the pandemic. Time-series analysis of the distribution of tweets over topics was performed to explore how the discussion on each topic changed over time, and the potential reasons behind the change. In addition, spatial analysis of topics was performed by comparing the percentage of tweets in each topic among top tweeting countries. Afterward, sentiment analysis of tweets was performed at both temporal and spatial levels. Our intention was to analyze how the sentiment differs between countries and in response to certain events. The performance of the topic model was assessed by being compared with other alternative topic modeling techniques. The topic coherence was measured for the different techniques while changing the number of topics. Results showed that the pooling by author before performing LDA significantly improved the produced topic models.

Keywords: COVID-19, Twitter, topic modeling, sentiment analysis, Latent Dirichlet Allocation, social media

Received: September 8, 2020 **Accepted:** December 11, 2020

***Corresponding Author:** Iyad AlAgha
https://orcid.org/0000-0002-7516-4902
E-mail: ialagha@iugaza.edu.ps

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far 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.

© Iyad AlAgha, 2021

1. INTRODUCTION

Starting in late 2019, the COVID-19 pandemic has rapidly impacted over 200 countries and territories (World Health Organization, 2020). It has caused an unprecedented impact on the world in ways that are not yet fully understood. Since the emergence of the disease, social media platforms, such as Twitter, have witnessed an extensive discourse on various aspects of COVID-19. This discourse reflects how individuals respond and are affected by the consequences of the disease. The analysis of the social media content related to COVID-19 can reveal patterns that can be vital for decision makers, health organizations, and businesses to respond intelligently in the current and future pandemics.

Twitter has been increasingly used by officials, politicians, and medical practitioners to disseminate information and communicate with people. The use of Twitter as a discussion platform on COVID-19 will only continue to grow, especially under the social distancing measures which have pushed people to spend more time on social media to keep updated and stay in touch. Thus, Twitter can serve as important data sources to provide collective intelligence and awareness of different aspects surrounding the disease, including health, economy, politics, and social life.

Numerous studies have utilized Twitter as a public perception monitor in different disastrous events, including tsunamis (Pourebrahim et al., 2019), earthquakes (Mendoza et al., 2019), hurricanes (Alam et al., 2018), and terrorist attacks (Bruns & Hanusch, 2017). In the context of disease outbreaks, Twitter has been also used as a data source for health surveillance and emergency communication systems (Byrd et al., 2016; Dai et al., 2017; Jordan et al., 2019). Twitter data has been found to be useful for several public health applications, including monitoring diseases, public reaction, outbreak/emergency situations, prediction, and lifestyle (Jordan et al., 2019).

This study evaluates the topics and opinions of COVID-19 discussion found on Twitter by using spatiotemporal topic modeling and sentiment analysis. Topic modeling will be used to determine the topics of discussion on Twitter about the pandemic. Sentiment analysis will be used to determine the different levels of positive and negative opinion on COVID-19 present in the dataset. Identifying these topics and sentiments will give decision makers a comprehensive view of the ongoing events and their impacts. It will also help them analyze and improve their policies based on the trends and moods perceived from Twitter.

Our work contributes to the growing body of literature on social media analytics and COVID-19. Compared to other studies related to COVID-19, this study is distinguished by the in-depth analysis that is performed at different levels (volume analysis, topic modeling, sentiment analysis) and from different perspectives (spatial, temporal). In addition, it employs and evaluates an enhanced topic modeling approach that uses Latent Dirichlet Allocation (LDA) with author-based pooling and phrase modeling to improve the quality of topics.

The rest of the paper is organized as the following: The following section discusses related works and highlights the differences between this study and existing ones. Afterwards, the experimental methodology used in this study is presented, followed by data collection and preparation steps. The topic modeling approach is then explained. Subsequently, volume analysis is performed to establish basic geospatial facts about the dataset. Then, results from topic modeling are presented from both temporal and spatial perspectives, and are discussed in light of significant real-world events where possible. Results from the sentiment analysis are then discussed. The performance of the topic modeling approach is then evaluated by being compared with other alternative techniques. The last section provides the summary and conclusions.

2. RELATED WORKS

Recently, a growing number of works have tried to analyze Twitter activity during the COVID-19 pandemic. These efforts vary based on their intentions, analysis techniques, and depth of analysis. Table 1 summarizes the key features of existing studies so far and compares them with our study. Some works (Banda et al., 2020; Kabir & Madria, 2020; Lopez et al., 2020) presented descriptive statistics and key features of Twitter datasets they collected, but they did not employ machine learning techniques to analyze discussed topics or sentiments. Cinelli et al. (2020) performed a comparative analysis on five social media platforms (Twitter, Instagram, YouTube, Reddit, and Gab) for the purpose of identifying prominent themes of discussion.

Several studies used topic modeling to determine the topics of discussion on Twitter about COVID-19. Singh et al. (2020) analyzed themes of discussion with a focus on myths and misinformation shared about the virus. Sha et al. (2020) analyzed tweets posted by the U.S. governors, the president, and his cabinet members. The intention was to extract topics that attracted attention, and to un-

 https://doi.org/10.1633/JISTaP.2021.9.1.3

Table 1. Comparison of studies that analyzed Twitter activity during COVID-19 pandemic

Author	Number of tweets	Period	Topic modeling	Sentiment analysis	Temporal analysis	Geospatial analysis	Network modeling
Sha et al. (2020)	7,881	Feb. 1 - Apr. 1, 2020	✓ (Hawkes Binomial Topic Model)		✓		✓
Ordun et al. (2020)	23,830,322	Mar. 24 - Apr. 9, 2020	✓ (LDA)	✓	✓		✓
Abd-Alrazaq et al. (2020)	1,150,825	Feb. 2 - Mar. 15, 2020		✓			
Wicke and Bolognesi (2020)	203,756	Feb. 4 – Mar. 20, 2020	✓ (LDA)				
Duong et al. (2020)	1,873,022	Jan. 20 - Mar. 20, 2020	✓ (LDA)		✓		
Medford et al. (2020)	126,049	Jan. 14 - Jan. 28, 2020	✓ (LDA)	✓	✓		
Kabir and Madria (2020)	100,000,000	Mar. 5 - Apr. 24, 2020					
Lopez et al. (2020)	6,468,526	Jan. 22 - Mar. 13, 2020				✓	
Singh et al. (2020)	2,792,513	Jan. 16- Mar. 15, 2020	✓ (Identify frequent words only)		✓	✓	
Cinelli et al. (2020)	1,187,482	Jan. 27 - Feb. 14, 2020	✓ (Word Embedding)				
Chen et al. (2020)	8,919,411	Mar. 5 - Mar. 12, 2020	✓ (LDA)				
Sharma et al. (2020)	30,800,000	Mar. 1 - Mar. 30, 2020	✓ (Word Embedding)	✓	✓	✓	
Banda et al. (2020)	30,990,645	Jan. 1 - Apr. 4, 2020					
Ours	7,516,668	Feb. 15 - May 31, 2020	✓ (LDA with author-pooling scheme)	✓	✓	✓	

LDA, Latent Dirichlet Allocation.

derstand how decision makers influenced each other by constructing influence networks. Ordun et al. (2020) used topic modeling to reveal common themes in Twitter data. They applied Uniform Manifold Approximation and Projection (UMAP) to visualize generated topics. However, user sentiment and geospatial analysis were not considered in their study. Abd-Alrazaq et al. (2020) performed topic modeling and then conducted sentiment analysis for each topic. However, their study did not include time based and geospatial analysis to compare the distribution of topics over time and across countries. Wicke and Bolognesi (2020) presented an analysis of the discourse around COVID-19 based on a corpus of 200k tweets. Their study used topic modelling, but it did not tackle sentiment, temporal, and special analysis of tweet contents. Duong et al. (2020) analyzed tweets from college students in terms of the response to the closure of universities and the discussed topics on several social issues. Medford et al. (2020) performed a sentiment analysis to identify emotional valence and predominant emotions during the COVID-19 outbreak. They also conducted topic modeling to explore discussion themes over time. Sharma et al. (2020) also used topic modeling to identify topics from Twitter posts on COVID-19. They also examined the sentiments over time in each country/region. Chen et al. (2020) presented a study on the topic preference related to the use of controversial and non-controversial terms associated with COVID-19 on Twitter.

To our knowledge, little work has been done so far to perform spatio-temporal analysis of topics and sentiments in COVID-19 Twitter content at a global level. Spatio-temporal analysis is essential to compare and contrast the nature of COVID-19 discussions between different countries and over time. In addition, the studies reported in Table 1 all cover early periods before the outbreak reaches its peak in top infected countries such as the USA and UK. This work inspects the time period from the middle of February to the end of May 2020. This period marked key moments in the timeline of the pandemic, such as reaching and surpassing the peak of the outbreak in certain countries. It also witnessed the tightening and easing of lockdown restrictions, which is likely to influence Twitter discussions and sentiments. Thus, our analysis is likely to reveal patterns and topics not observed

in the early stages of the outbreak.

Another technical difference between this study and the others lies in the technique used for topic modeling. While most existing studies used the standard LDA to infer topics from Twitter content (see Table 1), it has been found that the standard LDA does not work well with the messy form of Twitter content (Zhao et al., 2011). The naïve application of LDA to Twitter content is likely to produce incoherent topics (Mehrotra et al., 2013). Instead, this study uses the tweet pooling scheme from Hong and Davison (2010) to merge related tweets and present them as a single document to the LDA model. Evaluation of the topic coherence has proved improved results when compared to the topic model generated by the standard LDA.

3. EXPERIMENTAL METHODOLOGY

Fig. 1 depicts the experimental methodology adopted in this work. Several steps were performed sequentially in the workflow, including data collection, data preparation, volume analysis, topic modeling, sentiment analysis, and evaluation. Volume analysis aims to establish an overview on the geospatial features of the collected corpus. Topic modeling was then performed to identify the main topics of discussion. Topics were also analyzed over time and between top tweeting countries to provide an insight into the location-based and time-based differences. Sentiment analysis was also performed at both temporal and spatial levels, and results from the topic modeling were linked with the sentiment results where possible. Implications of each method of analysis were also discussed and put into context.

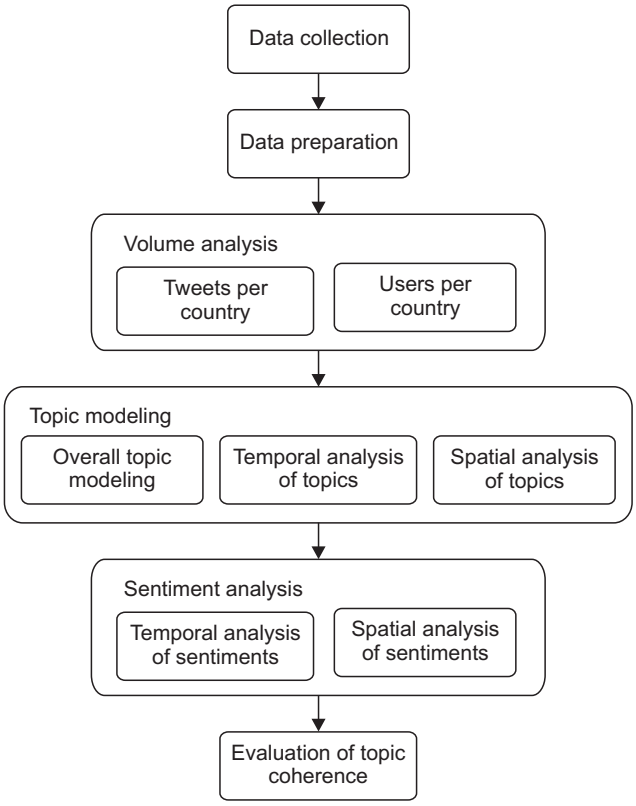


Fig. 1. Experimental methodology.

4. DATA COLLECTION

The dataset consists of 9,716,668 distinct tweets posted during the period from 15 February 2020 to 31 May 2020. It was collected by using Twitter application programmable interface (API). The keywords used to search for tweets were: coronavirus, COVID, COVID-19, ncov19, and ncov2019. The collected tweets should contain at least one of the aforementioned keywords or some variant of them. The following variables were collected for each tweet where possible: tweetid (an id associated with the tweet), userid (an id identifying the user who posted the tweet), postdate (the time and day the tweet was posted), latitude, longitude, and message (the text of the tweet). Note that the latitude and longitude variables may have null values if the Twitter API cannot identify the location where the tweet was posted from. IDs of all tweets in the dataset were made available online and can be accessed from https://github.com/iyadalagha/COVID-19-corpus (tweet information can be retrieved by the tweet id using Twitter API). Table 2 shows sample tweets and the associated information from the dataset. Note that the userids are not shown for privacy issues.

5. DATA PREPARATION

Tweets often have unique features that make the preprocessing of them different from that for ordinary text. Tweets are of limited length (140 characters at most), and may contain special texts such as hashtags, URLs, emoticons, and usernames. Tweets were pre-processed and prepared for the topic modeling and sentiment analysis by applying the following steps:

- Filtering: This study is limited to tweets written in English only, so all non-English tweets were excluded. In total, about 16% of collected tweets were written in non-English languages. In addition, all non-geocoded tweets were eliminated: These are the tweets that have no associated latitude and longi-

 https://doi.org/10.1633/JISTaP.2021.9.1.3

Table 2. Sample of tweets in the dataset

Message	Postdate	Latitude	Longitude	Country
2019: Soon, Amazon will use drones to deliver toilet paper! 2020: No toilet paper. No drones	3:02 PM - Mar. 14, 2020	36.778259	-119.417931	USA
Internet, social and webcam calls are becoming even more important in these days of isolation due to coronavirus. Here in North Italy we are adapting for online lessons at scuolacomicsbrescia and i see more and more…	Mar. 13, 2020	45.464664	9.188540	Italy
Australia has had 54 coronavirus deaths. To put this into context, on average 54 people die every 3 hours in Australia. The lockdown has worked. Now let's leave our borders closed and start re opening the economy.	5:02 AM - Apr. 12, 2020	-37.840935	144.946457	Australia
Testing for coronavirus AND treatment needs to be free. Period.	4:29 PM - Mar. 30, 2020	37.778259	-119.814931	USA
Report today on what's being done to protect minority ethnic NHS staff amid growing evidence around the disproportionate impact of coronavirus.	11:14 PM - May 4, 2020	53.234444	-0.538611	UK

NHS, National Health Service.

tude values. About 21.5% of the collected were non-geocoded.

- Tokenization and Tagging: Tweets often come with a special grammar and abbreviations so that users can convey messages with the least possible words. Traditional tokenizers and part of speech taggers may be inadequate for pre-processing tweets, and there is a need for alternatives that can recognize tweets' tokens, hashtags, emoticons, and URLs. We used a text processing library called ArkTweetNLP (Owoputi et al., 2013) to tokenize and tag tweets. The library was developed specifically to handle informal and online conversational text including various non-standard lexical items and syntactic patterns.
- Cleaning: Twitter users often use symbols and non-standard language in their tweets. Many of the used symbols may be irrelevant and thus may lead to incorrect results when analyzed. Thus, tweets were cleaned by removing the following parts: Usernames, numeric expressions, punctuation, and URLs. These parts were recognized from the tagger applied in the previous step.
- Normalization of emoticons: Emoticons are important for sentiment analysis; thus, their meanings should be preserved and should not be removed from tweets. In our approach, we used a special dictionary that contains the most used emotions and their meanings in English (Gimpel et al., 2013). This dictionary was used to replace each emoticon with its relevant meaning. For example, the emoticons ☺, ☺, and :) are all replaced with the word "happy."
- Spell check and correction: Tweets may contain incorrect or miss-spelled words. This step manipulates these words by using a spell checker, and substitutes them with correct words. Jazzy Spell Checker (Idzelis, 2005) was used for this step.
- Reverse geocoding: Tweets that have latitude and longitude values should be mapped to a readable country name. This is necessary to perform the geospatial analysis. The Google maps geocoding API was used to determine the country that corresponds to the input latitude and longitude. After assigning tweets to countries, tweets were counted per country. Countries that ended up with a number of tweets less than 0.1% of the total number of tweets were ignored.
- Stop-word and keyword removal: Stop-words were filtered out to improve the topic quality and reduce the noise. In addition, the keywords that were used to search for the tweets were also removed. Since these keywords are likely to be included in every tweet, removing them will ensure that the topics produced by topic modeling will not be dominated by the same keywords.
- Phrase modeling: One limitation of LDA is that it generates topics consisting of single words only. As a result, generated words that are originally part of frequently used phrases will become meaningless in results. For example, phrases such as "Stay at home" and "social distancing" will be split and presented as separate words after applying LDA. To resolve this issue, frequently used phrases in the corpus were

first detected by using the Gensim library in Python (Řehůřek & Sojka, 2011). For simplicity, only phrases of two and three words were detected if they occurred together *n* number of times, where *n* is a threshold value set to 100 based on our experiments. Each detected phrase is then put together into one word by replacing spaces with underscores (e.g., the phrase "Stay at home" becomes "Stay_at_home"). Hence it will be treated as a single word when processed by LDA.

After applying the aforementioned data preparation steps, the size of the final cleaned corpus was 6,058,710 tweets.

6. TOPIC MODELING USING LDA

Topic modeling is an unsupervised machine learning technique used to discover topics that occur within a collection of documents, and in what quantities (Blei & Lafferty, 2006). The topics generated by topic modeling are probabilistic mixtures of words that represent the word co-occurrence trends in the corpus. With respect to topic modeling in Twitter, each tweet is often treated as a document in the corpus, and can be viewed as a set of topics with different proportions depending on the frequency that terms appear (Alvarez-Melis & Saveski, 2016). Topics are the themes that the documents discuss or talk about. A good topic model is one that corresponds to distinct subjects of discussion in the dataset. This means that a good topic model should produce distinctive topics in the sense that the keywords belonging to a particular topic should not have much overlap with keywords in other topics.

LDA is a popular algorithm that is commonly used to train a topic model (Blei et al., 2003). Formally, LDA takes two inputs: 1) A bag-of-words corpus containing *d* documents and a vocabulary of size *v*, and 2) A scalar value *k*, which indicates the number of topics that the model will learn. LDA then outputs a model, *m*. The model *m* consists of two matrices:

1. A Topic × Vocabulary matrix, $T^m \in R^{k \times v}$. This is the matrix of topics that are learned by the model, with entry $T^m{}_{i,j}$ representing the association of word *j* with topic *i*.
2. A Document × Topic matrix, $D_m \in R^{d \times k}$, with entry $D^m{}_{i,j}$ representing the probability that document *i* is generated by topic *j*.

LDA works by looking at the word co-occurrences within documents. It then assumes that words that appear in the same document are likely to belong to the same topic, and that the documents that contain the same words are likely to contain the same topics. To summarize how the topics are inferred and to train an LDA model from the documents, the generative process is described as the following (Ponweiser, 2012):

1. For each topic, decide what words are likely.
2. For each document:
   a. Decide the proportions of topics that should be in the document.
   b. For each word:
      i. Choose a topic.
      ii. Given this topic, choose a likely word from 1.

LDA is often used on long documents. However, it has been found that it does not work well with shorter documents such as tweets and will often generate incoherent topics (Liu et al., 2018; Mehrotra et al., 2013). A common solution to this problem is tweet pooling (Alvarez-Melis & Saveski, 2016; Mehrotra et al., 2013), in which related tweets are merged together and presented as a single document to the LDA model. Several pooling schemes have been proposed in the literature to improve the LDA topic quality (Li et al., 2016; Mehrotra et al., 2013). These schemes include temporal pooling (tweets posted within the same hour), pooling by author, and pooling by hashtag. This work uses author-based pooling by building a document per author that includes all tweets he/she posted. Aggregating tweets according to the author is shown to be superior to un-pooled tweets with respect to topic modeling (Hong & Davison, 2010; Weng et al., 2010).

Another essential parameter of LDA is the number of topics that the model should produce from the corpus, *k*. Running the LDA model with *k* as input will result in dividing the documents in the corpus into *k* topics. Setting *k* to be too small will lead to too generic results because many distinct topics may be merged together. In contrast, setting *k* to be too big will result in too many topics that are individually confusing (Nikita, 2016). To choose the optimal number of topics *k*, the LDA model was built several times by setting *k*=5, 10, 15... and 50 (increments of 5), and the quality of the generated topics are compared. This approach was adopted in many works that used LDA for topic modeling (Maier et al., 2018; Zhao et al., 2015).

To address the memory and processing power needed to process the large volume of collected tweets and associ-

 https://doi.org/10.1633/JISTaP.2021.9.1.3

ated information, a cluster of computers was used in this study, and was operated by Apache Spark (Zaharia et al., 2016) as a cluster-computing framework. Apache Spark can distribute computational power to automatically parallelize and execute tasks on a large cluster of computers. It also provides a highly-optimized machine learning library called MLlib (Meng et al., 2016) which can perform LDA on numerous datasets.

7. RESULTS

7.1. Volume Analysis

Analyzing the volume of tweets in terms of posting countries and users is an essential first step in exploring the data and may provide some useful results by itself. It can also help us to better understand and put into context the results of topic modeling and sentiment analysis. Fig. 2 shows the number of tweets per country. The top five tweeting countries are the USA with 1,708,556 tweets, the UK with 690,693 tweets, Canada with 442,286 tweets, Australia with 387,757, and New Zealand with 254,466. The USA is the top tweeting country with respect to COVID-19, and contributes with 28.2% of the total number of tweets. Results also indicate that the countries that have the largest numbers of English-speaking populations are the most predominant in the corpus.

Another variable of interest is the number of users per country. In total, tweets in the corpus were posted by 1,553,516 users, and the average number of tweets per user is 4.9. Fig. 3 depicts the number of users per country. The distribution of users over countries is highly consistent with the distribution of tweets over countries. Over a

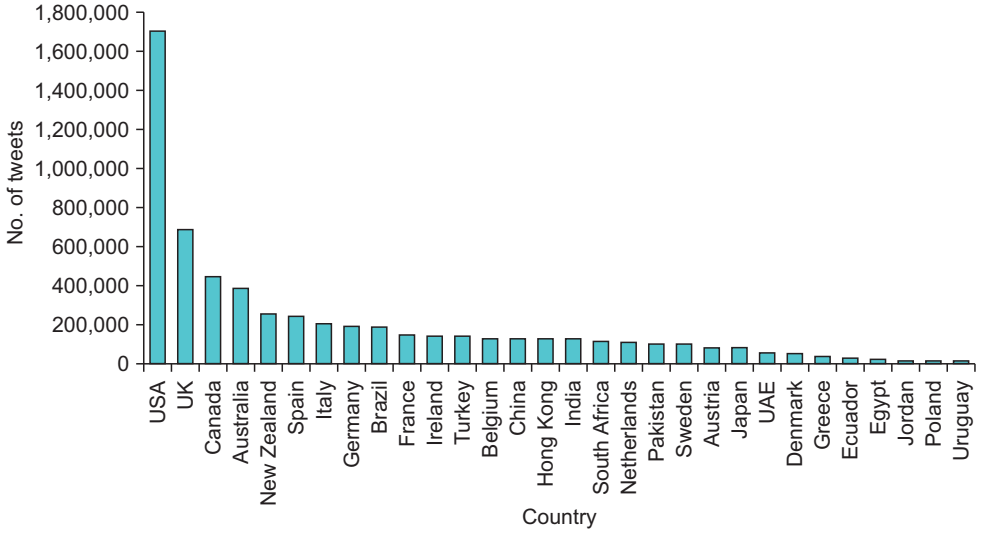


Fig. 2. Number of tweets per country.

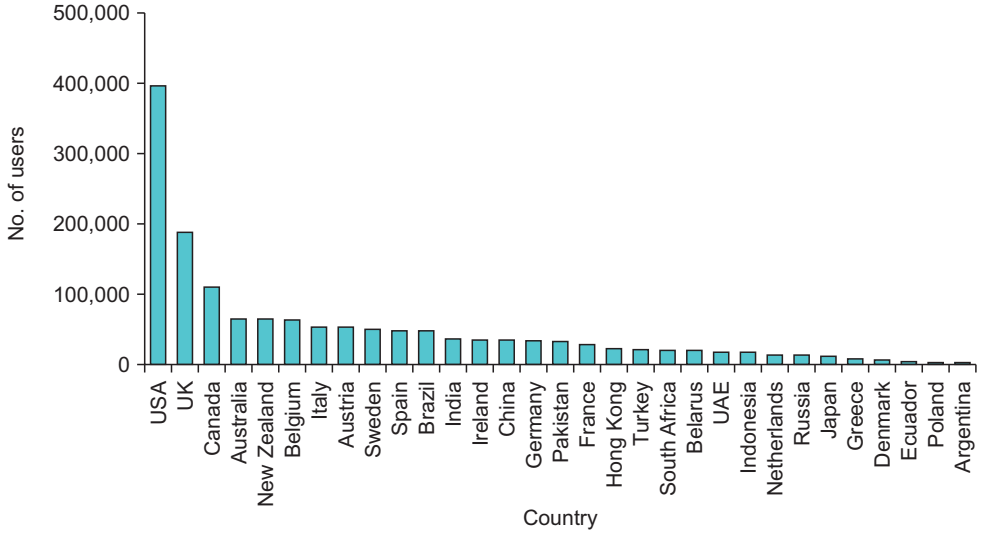


Fig. 3. Number of users per country.

quarter of the users in the corpus came from the USA. It is obvious from these results that the USA is the dominant country in the dataset in terms of the number of tweets and users. This is understandable given that the USA was the most affected country by the pandemic, during the period of study, as indicated by death and infection rates. Thus, it is likely that the attitudes and behavior of the USA users will have the highest influence on the topic modeling and sentiment analysis.

7.2. Topic Modeling

When testing LDA with different number of topics as an input parameter, results showed that using 20 topics resulted in the best topic model for the dataset. In addition, choosing 20 topics is a reasonable number that allows each topic to include a variety of narrower subjects.

7.2.1. LDA with author pooling scheme

Table 3 shows the results of topic modeling with topic labels. In general, words produced within several topics were mostly related enough to be given a label that describes the enclosing topic. The labels of topics were chosen by looking at the top 30 words per topic based on their probabilities. The ability to label the produced topics could be a sign of a good topic model. However, some topics look incomprehensible from a human perspective, and thus could not be labelled correctly. Fig. 4 visualizes the same topic model with word clouds, whereas the size of a word within a word cloud is proportional to the probability of the word within the topic.

After generating the topic model, each tweet in the dataset was assigned to its topic. This is necessary to understand the context in which words of the topic appear. LDA generates a document-topic matrix, which allows mapping tweets to the generated topics. Note that a single tweet can map to more than one topic. Thus, we mapped the tweet to the most probable topic so that topics are isolated for further analysis.

One should note that some generated topics may include incompatible words, and thus could not be labelled. As LDA starts with a degree of randomness, it generates a slightly different topic model every time, and the quality of some topics may vary largely. However, tuning LDA by setting the number of generated topics to 20 and by employing the author pooling scheme helped in stabilizing the generated topic model to a large extent in most iterations.

In addition, LDA does not discard words. It must assign each word to some topic, even if the word has relatively little meaning. This may cause some topics to be less coherent and meaningful. For example, the words "likes" and "doesn't" appear in many topics. However, many of the topics in Table 3 are indeed meaningful, which was the intended result.

It is normal in topic modeling to find many words appearing in several topics but with different weights. For example, the words "China," "Trump," "Stay_at_home," and "tests" appear in many topics, and this is a rational result when considering the overlap between the topics. In addition, several topics include multiword phrases such as "stay_home," "social_distancing," "wet_markets," "wash_hands," and "death_toll." These phrases appeared due to the phrase detection and grouping step applied before performing topic modelling. Without such a step, LDA will only generate topics of single words that may not convey precise meaning as compared to phrases.

It can also be noticed that some topics look similar, with small differences that may not be observable. For example, the topics "lockdown policies" and "protection" have many similar words. However, the difference between topics has a basis in the word co-occurrence statistics in the dataset. For example, the topic "lockdown policies" includes words that have frequently occurred in the context of governmental/national actions and plans in response to the COVID-19 outbreak. On the other hand, the words in the "protection" topic were mostly inferred from tweets discussing protection, safety products, and procedures. Similarly, the "Origin of COVID-19," "politics," and "economy" topics overlap in several words, but each topic is inferred from a different type of discussions.

The "politics" topic includes words that are mostly used in the context of government policies and international relations pertaining to COVID-19. Most tweets in the "politics" topic came from politicians and news agencies reporting updates related to the outbreak. In addition, the controversy raised over President Trump's tweets and White House briefings triggered political debates among USA users in particular. Numerous Twitter discussions that contributed to this topic also centered on tensions surrounding the trade and transport of medicines, diagnostic tests, and hospital equipment. Political figures like Donald Trump, Anthony Fauci, and Andrew Cuomo have gained attention in these discussions.

The economy is also strongly present in the topic model. The fears of economic fallout and job losses have been prevailing among Twitter users, and this was evident from the many topical words referring to the economic impact of COVID-19. Examples of these words are "Opening,"

 https://doi.org/10.1633/JISTaP.2021.9.1.3

Table 3. Top words of 20 topics generated from LDA with author-pooling scheme

Prayers and blessings: stay at home, home, stay_ home, bless, safe, vaccine, god, well, keep, testing, face, distance, information, clean, help, negative, tests, said, cdc, health, keep, infectious, hospital, donation, doctors, like, please, positive, survive, birx, get, protection, survived, players, project	**Lockdown policies:** stay_at_home, social_ distancing, distance, hands, information, sanitizer, mask, wash_ hands, openning, response, who, lockdown, safe, cdc, help, flights, reopening, virus, closure, schools, country, emergency, pandemic, responses, universities, fauci, need, webcam, governor, paid, fund, infection, economy, government, relief, act, aid	**Origin of COVID-19:** china, animal, animals, trump, wuhan, infection, chinese_virus, spread, hospital, virus, wet_markets, who, cases, travel, wet, bat, deaths, disease, lab, infected, world, lockdown, bioweapons, people, markets, tiktok, infectious, time, got, said, lol, facts, like, vaccine, tested, health	**Symptoms:** cough, flu, dry, high, fever, temperature, allergy, headache, doctors, help, hospital, breathing, testing, rash, loss, toilet_paper, infectious, nose, medical, test, smell, heat, cdc, heat, vaccine, infection, public, protect, arkansas, seek, 5g, dont, tested, new, today	**Protection:** social_distancing, tests, sanitizer, alcohol, wash_hands, outbreak, masks, care, advice, vaccine, toilet_paper, protection, medical, treatment, drug, keep, who, donald, clean, immune_system, policy, gloves, positive, virus, chloroquine, get, good, cdc, protect, likes, low, national, ohio, said, need
Healthcare: medical, nurses, cases, hospital, doctors, test, patients, ventilators, deaths, beds, shortage, emergency, country, negative, nhs, positive, trump, diseases, us, vaccine, scientists, chloroquine, spread, cases, cure, national, tests, online, public, use, died, donation, due, research, human	**Politics:** trump, vaccine, china, lockdown, tests, health, president, wuhan, borders, press, injecting, disinfectant, state, cnn, pence, pence, chinese_ virus, said, spread, world, climate, plan, aid, nyc, boris, new, report, says, science, shortage, strategy, bioweapon, inject, confinement, fauci, lol, opening, lose, problems, response, going, racist	**Economy:** president, china, opening, online, recession, bank, jobs, stock, face, gold, industry, economic_ recession, lockdown, graph, economy, emergency, prices, world, amazon, billions, fake, act, money, tourism, crisis, impact, plan, welfare, nyc, businesses, climate, diamond, paid, due, god, relief, says, going, shortage, benefits, response, suspended	**Infected countries:** death_toll, china, increase, patients, deaths, us, italy, spread, updates, wuhan, cases, country, positive, online, uk, spain, pandemic, public, brazil, nyc, latest, rates, graph, cuomo, good, recovery, taste, new jersey, cases, fake, dont, racist, france, germany, update, new york, illinois, climate	**Lockdown activities:** flights, increase, online, tests, immune_system, movie, netflix, people, football, hair, public, tiktok, series, heat, shopping, smell, need, wuhan, fun, stock, gaming, quarantine, positive, meeting, stay_home, like, likes, economy, ship, fauci, suspended, confinement, lol, 5g, low, makes, protect, sport
toilet_paper, food, flights, delivery, sanitizer, masks, fitness, travel, gaming, music, webcam, online, lol, soap, hair, hand, netflix, increase, camera, diapers, like, gloves, taste, zoom, public, nyc, stores, shopping, know, movie, ca, science, shortage, racism	donald, china, trump, disease, home, vaccine, italy, lol, gaming, smell, heat, population, music, health, virus, industry, emergency, news, flights, reopen, delivery, action, like, 5g, universities, infected, economy, social, zoom, keep, plan, science, change, masks, decline, market, racist	chinese, disease, spread, travel, virus, home, loss, hospital, lockdown, tests, said, protect, boris, fever, infected, people, vaccine, amazon, early, breath, stores, nhs, cuomo, hands, nyc, increased, governor, mild, lockdown, doctors, food, china, infection, germany, racist, symptoms	**COVID-19 in the USA:** home, states, disease, hospital, news, nyc, president, trump, washington, cnn, fauci, state, new york, healthcare, industry, spread, tests, high, nyc, cdc, recession, new jersey, ca, january, jobs, lockdown, health, february, food, cuomo, infection, spreading, city, birx, arkansas, government, like, says, large	medical, china, hospital, patient, vaccine, stay_home, confinement, cough, flu, online, spread, cases, symptoms, wet_markets, travel, closure, pandemic, racist, tested, nhs, ventilators, cost, day, cuomo, universities, travel, economy, positive, policy, says, world, social, call, hey, fitness, think, timeline
beds, event, hospital, economy, mask, vaccine, closure, stock, stores, music, closure, economic_recession, change, home, sanitizer, good, cost, netflix, face, flights, infectious, please, industry, travel, said, public, says, work, human, bat, markets, government, pray, scientist, hands, reduce, sad	masks, hospitals, universities, billion, lack, 5g, protection, safety, loss, fake, doctors, nurse, birx, closure, public, research, vaccine, sanitizer, america, updates, nyc, climate, drop, like, problems, lab, rules, world, spain, support, claims, winter, breath, save, symptom, warning	china, medical, doctors, information, infection, said, wuhan, economy, flu, president, tiktok, heat, markets, bat, vaccine, died, birx, hospitals, march, uk, said, social_distancing, going, day, bioweapons, tested, public, reopen, spread, wet, spreading, help, said, sport, work	health, chinese, says, pandemic, us, hospital, organization, infectious, health, like, beds, event, home, bat, protect, closure, sanitizer, breath, aid, china, cnn, control, day, faces, facts, home, informed, like, news, online, policy, public, safe, vaccine, ventilators, wash	doctors, disease, america, china, world, 5g, wuhan, health, job, economy, tests, positive, press, bioweapons, birx, lockdown, virus, immune, racist, says, information, cases, events, great, loss, sanitizer, called, increase, information, issued, jobs, plan, protect, tests, threat

LDA, Latent Dirichlet Allocation.



Fig. 4. Word clouds for Latent Dirichlet Allocation topic model.

 https://doi.org/10.1633/JISTaP.2021.9.1.3

"bank," "jobs," "stock," "recession," "economic_recession," "welfare," "prices," "industry," and "emergency."

It was also noticed that there are topics that reflect the top information needs and seeking behavior of Twitter users during the COVID-19 outbreak. Topics labelled as "symptoms," "protection," and "the origin of COVID-19" collectively show that the discovery of and protection from the disease were among the main topics raised in Twitter conversations. Another remarkable observation is the presence of a topic on healthcare. Words in this topic were mainly inferred from tweets discussing the concerns and difficulties faced by medical staff in infected countries and the shortage in hospital beds, ventilators, and medical staff.

Another interesting result is the presence of several words that indicate how people behave or spend their time during the quarantine. One topic is labelled as "Lockdown activities" since it includes words such as "movie," "series," "music," "toilet_paper," "gaming," "wash_hands," "webcam," and "hair." In fact, these words refer to products or activities that flourished during the lockdown. In addition, the topic model includes the names of several products that experienced high demand and spikes in sales such as "alcohol," "hand sanitizer," "masks," and "ventilators." The analysis of tweets related to this topic showed the increase in discussion around the closure of stores, and the shortage of certain products in several countries.

It is also interesting to find names of certain businesses such as "Amazon," "TikTok," and "Netflix" repeated in multiple topics. This result makes sense considering that the services offered by these businesses have seen an enormous increase in demand during the quarantine period. The word "online" also appeared repeatedly, indicating that consumers are shifting towards online activities in daily life. In general, these results shed light onto consumer behavior and spending habits during the pandemic.

Among the generated topics, there is also a topic dedicated to debate over the origins of COVID-19. This particular topic includes words that link to several conspiracy theories that were observed over multiple countries, such as the conspiracy promoting that COVID-19 is a bioweapon, the correlation between 5G technology and the virus, and the role of Chinese labs and bats. These terms, which also occurred in different other topics, highlight the widespread propagation of false information that has misled or influenced public opinion. Twitter discussions related to this topic also reveal the increased use of controversial terms insinuating that China is the origin of the virus. However, the controversy over the origin of the virus was prominent during March 2020 when many public figures used terms such as "the Chinese virus" in their tweets to refer to COVID-19, but it then subsided in April and May, especially after Twitter started labelling misleading, disputed, or unverified tweets about the coronavirus.[1]

7.2.2. Temporal analysis

A temporal analysis of the topic model was performed. The purpose was to explore how the discussion on each topic changed over time, and the potential reasons behind the change. For simplicity, some related topics were examined as a group. For example, Fig. 5 shows a time series plot of the number of daily tweets belonging to the topics "politics" and "economy." The number of daily tweets was generally increasing from the first of March till the first of

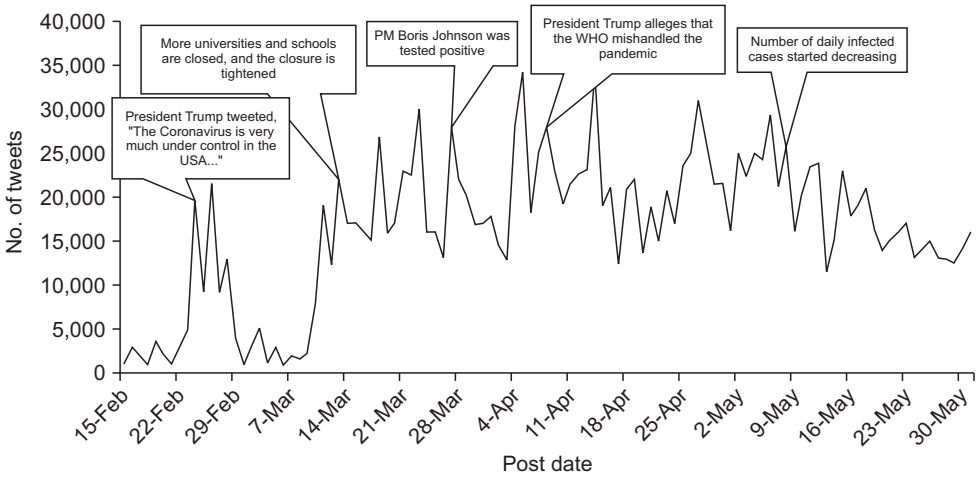


Fig. 5. Proportion of "politics" and "economy" topic discussion per day. PM, Prime Minister; WHO, World Health Organization.

[1]https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

May. The major increase was around the middle of March when nationwide social distancing protocols and school closures were attracting mass concerns. Afterwards, the daily number of tweets continued to increase at the same rate, before declining in May when most countries started to pass the peak of the outbreak and ease lockdowns. Thus, this pattern is in correspondence with the timeline of the infection and the changes in lockdown rules in the USA and the EU.

The spikes in the discussion on "politics" and "economy" topics were also explored and linked to certain events where possible (see data labels in Fig. 5). It was noticed that many spikes coincide with the discussion over the tweets of President Trump or lockdown procedures in the USA. In general, the discussion pattern was mostly influenced by political events and the timeline of the outbreak in the USA in particular. This indicates that politics/economy topics have a bias towards the USA as compared to other countries. This bias can be explained by the larger number of tweets in politics/economy that came from the USA. About 40% of USA tweets were in the politics/economy topics as compared to 30% of UK tweets and 22% of Canada tweets.

Fig. 6 shows a similar time-series plot of the number of tweets belonging to the topics "symptoms," "protection," and "healthcare." The discussion over these topics started to rise after the first week of March, before reaching a peak in April, and then declining in May. Again, this is consistent with the timeline of the outbreak: In March, discussions over protection and lockdown protocols developed considerably in response to the outbreak and the pressure on healthcare services. The discussions then started to drop in May after most countries reached or passed the peak of infections.

7.2.3. Spatial analysis

In the previous analysis, topic modelling was performed at the world level by running LDA over the entire dataset. To explore the differences between individual countries, topic modeling was performed on certain English-speaking countries. LDA with an author pooling scheme and 20 topics was applied to the tweets of the USA, the UK, and Canada separately. Differences between generated topic models were then explored. Overall, the topic models of the UK and Canada were poor. The topics in the UK and Canada corpora were less meaningful and incoherent. Several topics were indistinguishable from each other, and many others seemed to cover multiple subjects at once. Only the model of the USA was the most similar to the model on the entire dataset, though the quality was lower. In general, the poor models produced for individual countries is likely caused by the division of the entire collection of tweets into smaller corpuses that were not sufficient for LDA to train accurate models. However, the larger size of the USA corpus compared to other countries resulted in a model with seemingly better quality.

Based on the above discussion, it was not feasible to run LDA on county corpuses and then calculate the proportion of each topic discussion at the level of each country. To resolve this issue, just one topic model is produced using the entire dataset, and the percentage of tweets in each topic per country was analyzed. This approach was carried out by using the topic model shown in Table 3. For each of the aforementioned English-speaking countries, tweets in major topics ("politics," "economy," "protection & symptoms," "healthcare," and "prayers and blessings") were separated, and the percentage of topic discussion per country was calculated. Fig. 7 shows these percent-

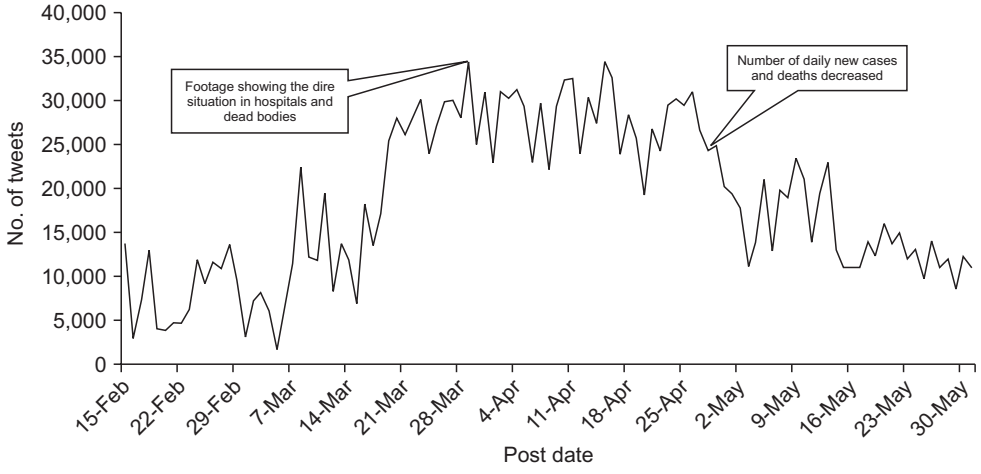


Fig. 6. Proportion of "symptoms" and "protection" and "healthcare" topic discussion per day.

 https://doi.org/10.1633/JISTaP.2021.9.1.3

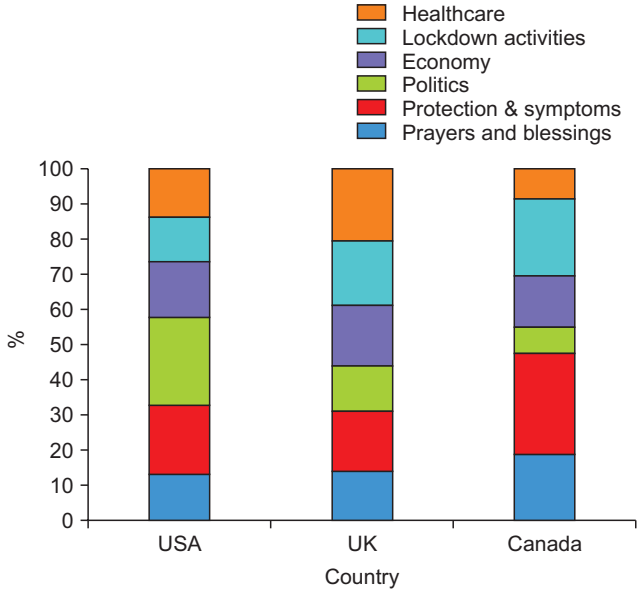


Fig. 7. Topic discussion for select topics by country.

ages. The percentage of topic discussion is defined as the number of tweets on that topic divided by the number of tweets on all topics times 100.

Notice that the "politics" topic made more discussion in the USA (25%) than in the UK (13%) and Canada (8%). This indicates that the tweets in the USA are more responsive to political events. This result is also consistent with the time-based analysis in Fig. 5, which shows that the growth in tweet numbers was influenced by certain political events. Discussion on the "economy" topic contributed by 15%, 17%, and 14% of the overall Twitter conversation in the USA, the UK, and Canada respectively. This indicates that the economic impact of COVID-19 was seemingly a shared concern in the three countries. The situation of healthcare staff and services also raised considerable discussion in the UK (21%), followed by the USA (14%) and Canada (9%). These numbers are apparently consistent with the severity of the outbreak and the pressure on the healthcare systems in the three countries. The little discussion on the healthcare topic in Canada can be explained by the fact that it is the least affected by the disease during the studied period.

Looking at the discussion on the "Lockdown activities" topic, it can be noticed that there was a significant Twitter discussion in the three countries about indoor activities that can alleviate the lockdown boredom. The analysis of this particular discussion can be of importance to business analyzers to better understand consumer behavior and preferences during the pandemic.

7.3. Sentiment Analysis

While volume analysis and topic modeling can be used to identify what topics are discussed in different countries or over a period of time, sentiment analysis is useful to determine emotional state in response to the outbreak.

To measure the sentiments of tweets, we used Valence Aware Dictionary and Sentiment Reasoner (VADER). VADER is a sentiment analysis model that was created for social media sentiment analysis (Hutto & Gilbert, 2014). It is more sensitive to sentiment expressions in social media contexts. The developers of VADER compared it with eleven state-of-the-practice benchmarks, and results showed that VADER improved the benefits of traditional sentiment lexicons (Hutto & Gilbert, 2014).

VADER was used to calculate a score for each tweet in the dataset. The calculated score ranges between -1 and +1, where -1 indicates the most negative and +1 indicates the most positive. Then, tweets were aggregated by post-dates and countries in order to perform temporal and geo-spatial analysis of sentiment. The aim is to analyze how the sentiment has changed in different countries and over time. In addition, we aim to see whether spikes in positive or negative sentiment can be associated with certain events.

Fig. 8 shows a time-series plot of the total sentiment per day over the entire world. In general, the period from 15 February to 11 April was dominated by negative discussions. Looking at the tweets posted in this period, it was obvious that the feelings of fear and anxiety were strongly present. In addition, intensive debates raged over the preparedness of healthcare services and the response plans of governments. These debates were often filled wih derogatory and ridiculous words. After the middle of April, the sentiment started to swing towards the positive direction. In particular, the announcement of reopening plans in many countries and the ease of restrictions made a positive impact. This positivity was also observed from the increasing discussion during May on individuals' plans to resume normal life activities.

It may be also interesting to link the change in sentiment in Fig. 8 with the distribution of discussed topics shown in Figs. 5 and 6. Major topics on politics, economy, symptoms, and protection were heavily discussed in March and April, before subsiding in May. Meanwhile, the average sentiment score was typically negative while the discussion on the aforementioned topics was active. Then, the sentiment score started to go positive as the infection started to decline and the discussion on the topics started to fade.

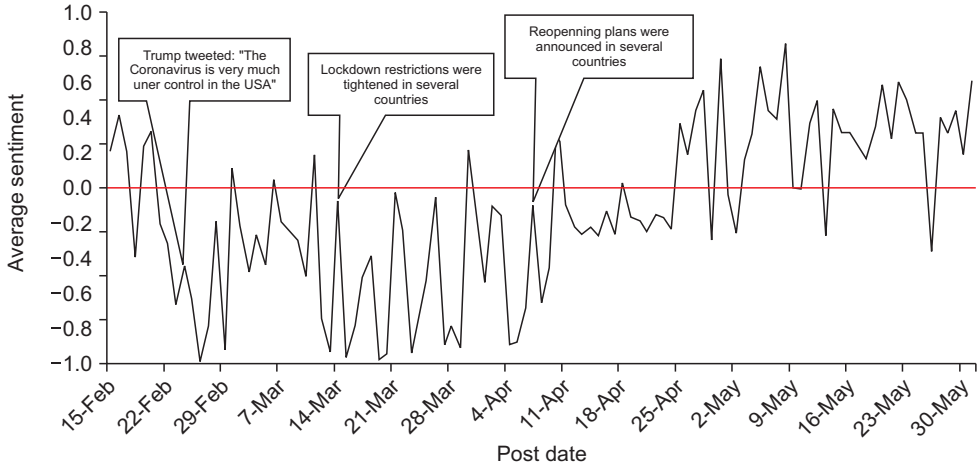


Fig. 8. Average sentiment of tweets per day.

Fig. 9 shows the average sentiment of all tweets per fortnight by the top three tweeting countries (the USA, the UK, and Canada). The average sentiment per country for the whole period (15 February to 31 May) is -0.0261 for the USA, -0.2279 for the UK, and 0.2053 for Canada. This result makes sense because Canada is the least affected country by the pandemic during the studied period. The sentiment of Canada remained positive for the whole period, ranging between 0.115 and 0.401. The USA and the UK almost had negative sentiments until late April-May when the sentiment started going up. The three countries shared the same peak around the middle of May.

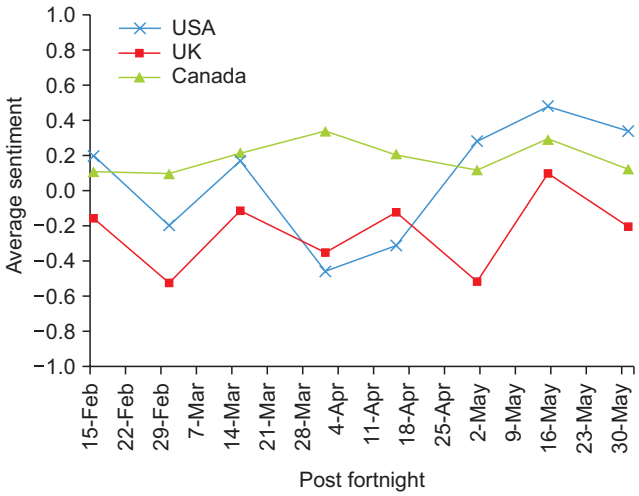


Fig. 9. Average sentiment of tweets per fortnight between the three top tweeting countries.

8. EVALUATION

In this part, we aim to assess the interpretability of the generated topic models. The evaluation of topic modeling has been addressed in the literature by two schools: The first school adopted a human-based assessment in which humans read topics to evaluate their interpretability (Aletras & Stevenson, 2014; Rossetti et al., 2016; Smith et al., 2017). The second school has taken a more principled approach by employing measures that can automatically judge the quality of topics (Aletras & Stevenson, 2013; Mehrotra et al., 2013; Newman et al., 2010). A class of these measures are topic coherence measures, which are used to score a topic by measuring the degree of semantic similarity between the top words in the topic. These measurements help distinguish between topics that are semantically interpretable topics, and topics that are artifacts of statistical inference (Stevens et al., 2012). For the evaluation in this study, we consider the UMass metric (Mimno et al., 2011). UMass computes the coherence of a topic as the sum of pairwise distributional similarity scores over the set of topic words as the following:

$$UMass(W_T) = \sum_{(w_i,w_j)\in W_T} sim(w_i,w_j,\in)$$

where W_T is the set of words in topic T. $sim(w_i,w_j,\in)$ is the distributional similarity between a pair of words w_i,w_j. $\in$ is a smoothing factor which guarantees that sim returns a real number ($\in$=1 is normally used). $sim(w_i,w_j,\in)$ is calculated as the following:

$$sim(w_i,w_j,\in) = log\frac{D(w_i,w_j)+\in}{D(w_j)}$$

where $D(w_i,w_j)$ counts the number of documents containing words w_i,w_j, and $D(w_j)$ counts the number of docu-

 https://doi.org/10.1633/JISTaP.2021.9.1.3

ments containing w_j.

A topic with a good UMass score contains words that are related to each other, i.e., words that are more likely to appear in the same tweets, whereas a topic with a bad UMass contains words that are unlikely to appear together in the same tweets. After calculating the UMass score of a topic, the performance of the whole topic model is calculated as the average of the UMass scores of its topics.

The quality of the model generated with LDA with author pooling scheme was compared with three other techniques commonly used in the literature for topic modeling. These techniques are:

- Standard LDA, in which each tweet is treated as a single document, and LDA is trained on all tweets. This serves as a baseline for comparison to other techniques.
- LDA with hashtag-based pooling: Hashtags are often used as topical markers for tweets. Users often use the same hashtags if they share related ideas or contribute with similar content. Of the 6,058,710 tweets in our dataset, 1,686,738 tweets (27.8%) have hashtags. Only tweets with hashtags can be used in this technique while other tweets are ignored. Tweets with the same hashtag were grouped as a single document. If a tweet had multiple hashtags, it was added to the document of each of those hashtags. This technique is used in Alvarez-Melis and Saveski (2016) and Sharma and Bedi (2018).
- LDA with temporal pooling: When a major event occurs, a large number of users often start tweeting about the event within a short period of time. In this technique, tweets posted within the same hour were gathered in a single document. Then, the LDA model was trained on the collection of generated documents. This technique is used in Mehrotra et al. (2013).

The aforementioned three techniques were compared against the LDA with author-pooling scheme used in this study. For each technique, the same pre-processing and data cleaning steps explained in the data preparation step were applied. For each technique, a topic model was produced while changing the number of topics from 5 to 50 with a step of 5. The aim was also to investigate the optimal number of topics for each technique with our dataset. Table 4 shows UMass scores of different topic modelling techniques.

Fig. 10 depicts how the average UMass score changes as the number of generated topics increases. It indicates that the performance of the topic models reaches a peak between 20 to 30 topics for all techniques. Then, the performance starts degrading as the number of topics exceeds 30. In particular, LDA with author pooling generated the best model with 20 topics.

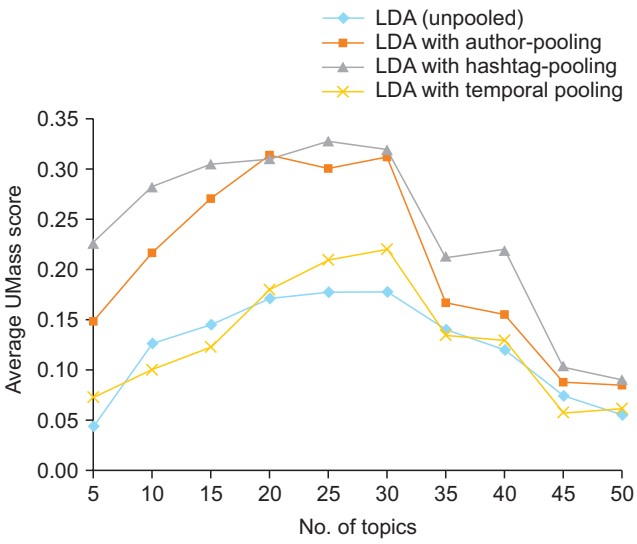


Fig. 10. Change in UMass score when number of topics changes for different topic modeling techniques. LDA, Latent Dirichlet Allocation.

Table 4. Comparison between topic modeling techniques based on UMass coherence score

Topic modeling technique	Average UMass score									
	5 topics	10 topics	15 topics	20 topics	25 topics	30 topics	35 topics	40 topics	45 topics	50 topics
LDA (un-pooled)	0.0448	0.1286	0.1453	0.1723	0.1782	0.1784	0.1404	0.1221	0.0747	0.0566
LDA with author-pooling	0.1486	0.2165	0.2711	0.3144	0.3018	0.3128	0.1668	0.1564	0.0877	0.0845
LDA with hashtag-pooling	0.2275	0.2841	0.3054	0.3106	0.3282	0.3201	0.2127	0.2203	0.1033	0.0901
LDA with temporal pooling	0.0744	0.1014	0.1233	0.1815	0.2101	0.2214	0.1366	0.1303	0.0583	0.0619

LDA, Latent Dirichlet Allocation.

Overall, LDA with hashtag pooling gives the best UMass score with 20 topics. LDA with author pooling comes second with a marginal difference. However, it outperforms the other two techniques, i.e., the standard LDA and the LDA with temporal pooling. Given these results, LDA with author pooling is still considered the best choice for this study since only 27.8% of tweets are associated with hashtags.

9. CONCLUSIONS AND FUTURE WORK

This work presents a study of Twitter discussion on the COVID-19 pandemic for the purpose of analyzing themes of discussion and sentiment. LDA with author pooling scheme was used to extract topics. Topics are then discussed to highlight significant observations and potential implications. Our study differs from existing analytical studies on COVID-19 in Twitter in the following aspects: 1) Topic modeling and sentiment analysis are performed from both temporal and spatial perspectives, which allowed in-depth insight into the impact of the pandemic on topics and sentiments over time and between countries. Results showed that both topics and sentiments are influenced by significant events taking place during the timeline of the outbreak. In addition, countries had different topics of interests and sentiments depending on the preparedness of each country and how much it was affected by the infection. The analysis also uncovered several patterns, such as the kind of lockdown activities that people practiced, the businesses and products that were in high demand, and the misinformation swirling around COVID-19. 2) The study uses LDA with author pooling scheme and phrase detection for topic modeling. This is in contrast to most other studies which used an un-pooled LDA. The intention was to overcome the limitations encountered with the standard LDA when used with short texts such as tweets. 3) To our knowledge, the study covers a more recent and longer period of time that includes significant events not covered in previous studies. During the studied period, the lockdown policies were tightened and eased, and the infection reached the peak in the top tweeting countries.

The study has the following limitations: 1) It only examined tweets in English, and thus the results cannot be generalized to non-English speaking countries. 2) Many tweets in the dataset are difficult to understand because the context of the tweet, i.e., the conversation, is not preserved. This can affect the quality of the topic modeling and sentiment analysis, and make it difficult to perform text mining.

There are many directions to extend this work. Since the pandemic is at a global scale, analysis of other social media platforms and different languages is critical to gain a better understanding of the global consequences of COVID-19. Second, it is possible to analyze the retweet activity to better understand how information is propagated and which users dominated the COVID-19 discussions. Third, network analysis can be performed to identify clusters of users with topic similarity. Fourth, it may be interesting to analyze the Twitter activity of specific groups of users such as the people who were tested positive, or people who are more vulnerable to COVID-19.

CONFLICTS OF INTEREST

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

REFERENCES

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top concerns of Tweeters during the COVID-19 pandemic: Infoveillance study. *Journal of Medical Internet Research*, 22(4), e19016. https://doi.org/10.2196/19016.

Alam, F., Ofli, F., Imran, M., & Aupetit, M. (2018). A Twitter tale of three hurricanes: Harvey, Irma, and Maria. *arXiv*. https://arxiv.org/abs/1805.05144.

Aletras, N., & Stevenson, M. (2013, March 19-22). Evaluating topic coherence using distributional semantics. In A. Koller & K. Erk (Eds.), *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* (pp. 13-22). Association for Computational Linguistics.

Aletras, N., & Stevenson, M. (2014, June 22-27). Labelling topics using unsupervised graph-based methods. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 631-636). Association for Computational Linguistics.

Alvarez-Melis, D., & Saveski, M. (2016, May 17-20). *Topic modeling in Twitter: Aggregating tweets by conversations*. Paper presented at the 10th International AAAI Conference on Web and Social Media, Cologne, Germany.

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E., & Chowell, G. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research -- an international collaboration. *arXiv*. https://

arxiv.org/abs/2004.03688.

Blei, D. M., & Lafferty, J. D. (2005, December 5-10). Correlated topic models. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18 (NIPS 2005)* (pp. 147-154). MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993-1022. https://doi.org/10.1145/2615569.2615680.

Bruns, A., & Hanusch, F. (2017). Conflict imagery in a connective environment: Audiovisual content on Twitter following the 2015/2016 terror attacks in Paris and Brussels. *Media, Culture & Society*, 39(8), 1122-1141. https://doi.org/10.1177%2F0163443717725574.

Byrd, K., Mansurov, A., & Baysal, O. (2016, May 14-15). Mining Twitter data for influenza detection and surveillance. In P. Kellenberger (Ed.), *2016 IEEE/ACM International Workshop on Software Engineering in Healthcare Systems (SEHS)* (pp. 43-49). Institute of Electrical and Electronics Engineers.

Chen, L., Lyu, H., Yang, T., Wang, Y., & Luo, J. (2020). In the eyes of the beholder: Analyzing social media use of neutral and controversial terms for COVID-19. *arXiv. https://arxiv.org/abs/2004.10225.*

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *arXiv*. https://arxiv.org/abs/2003.05004.

Dai, X., Bikdash, M., & Meyer, B. (2017, March 30-April 2). From social media to public health surveillance: Word embedding based clustering method for Twitter classification. In Institute of Electrical and Electronics Engineers (IEEE) (Ed.), *SoutheastCon 2017* (pp. 1-7). IEEE.

Duong, V., Pham, P., Yang, T., Wang, Y., & Luo, J. (2020). The ivory tower lost: How college students respond differently than the general public to the COVID-19 pandemic. *arXiv*. https://arxiv.org/abs/2004.09968.

Gimpel, K., Schneider, N., & O'Connor, B. (2013). *Annotation guidelines for Twitter part-of-speech tagging version 0.3 (March 2013).* http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf.

Hong, L., & Davison, B. D. (2010, July 25). Empirical study of topic modeling in Twitter. In P. Melville, J. Leskovec, & F. Provost (Eds.), *Proceedings of the first workshop on social media analytics* (pp. 80-88). ACM. https://doi.org/10.1145/1964858.1964870

Hutto, C. J., & Gilbert, E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. Paper presented at 8th International AAAI Conference on Weblogs and Social Media, Ann Arbor, Michigan, USA.

Idzelis, M. (2005). *Jazzy*: *The Java open source spell checker*. http://jazzy.sourceforge.net.

Jordan, S. E., Hovet, S. E., Fung, I. C.-H., Liang, H., Fu, K.-W., & Tse, Z. T. H. (2019). Using Twitter for public health surveillance from monitoring and prediction to public response. *Data*, 4(1), 6. https://doi.org/10.3390/data4010006.

Kabir, Y., & Madria, S. (2020). CoronaVis: A real-time COVID-19 Tweets data analyzer and data repository. *arXiv*. https://arxiv.org/abs/2004.13932.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016, July 17-21). Topic modeling for short texts with auxiliary word embeddings. In R. Perego & F. Sebastiani (Eds.), *SIGIR '16: The 39th International ACM SIGIR conference on research and development in Information Retrieval* (pp. 165-174). ACM.

Liu, C., Liu, Z., Li, T., & Xia, B. (2018, July 1-3). Topic modeling for noisy short texts with multiple relations. In X. He (Ed.), *Proceedings of the 30th International Conference on Software Engineering and Knowledge Engineering* (pp. 610-609). KSI Research Inc. and Knowledge Systems Institute Graduate School. http://ksiresearchorg.ipage.com/seke/seke18.html

Lopez, C. E., Vasu, M., & Gallemore, C. (2020). Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. *arXiv*. https://arxiv.org/abs/2003.10359.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3), 93-118. https://doi.org/10.1080/19312458.2018.1430754.

Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., & Lehmann, C. U. (2020). An "infodemic": Leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. *Open Forum Infectious Diseases*, 7(7), ofaa258. https://doi.org/10.1093/ofid/ofaa258.

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013, July 28-August 1). Improving LDA topic models for microblogs via Tweet pooling and automatic labeling. In G. J. F. Jones & P. Sheridan (Eds.), *SIGIR '13: The 36th International ACM SIGIR conference on research and development in information retrieval* (pp. 889-892). ACM. https://doi.org/10.1145/2484028.2484166

Mendoza, M., Poblete, B., & Valderrama, I. (2019). Nowcasting earthquake damages with Twitter. *EPJ Data Science*, 8, 3. https://doi.org/10.1140/epjds/s13688-019-0181-0.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S.,

Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M. J., Zadeh, R., Zaharia, M., & Talwalkar, A. (2016). MLlib: Machine learning in Apache Spark. *Journal of Machine Learning Research*, 17(2016), 1-7. https://www.jmlr.org/papers/volume17/15-237/15-237.pdf.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011, July 27-31). Optimizing semantic coherence in topic models. In P. Merlo (Ed.), *EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 262-272). Association for Computational Linguistics.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June 2-4). Automatic evaluation of topic coherence. In R. M. Kaplan (Ed.), *HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100-108). Association for Computational Linguistics.

Nikita, M. (2015). *Select number of topics for LDA model*. http://rstudio-pubs-static.s3.amazonaws.com/107657_4cdc6f600fe44cc8b2600f6f9c470ce8.html.

Ordun, C., Purushotham, S., & Raff, E. (2020). Exploratory analysis of Covid-19 Tweets using topic modeling, UMAP, and DiGraphs. *arXiv*. https://arxiv.org/abs/2005.03082.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013, June 9-14). Improved part-of-speech tagging for online conversational text with word clusters. *Proceedings of NAACL-HLT 2013* (pp. 380-390). Association for Computational Linguistics. https://www.aclweb.org/anthology/N13-1039.pdf

Ponweiser, M. (2012). *Latent Dirichlet allocation in R*. [Diploma thesis]. Vienna University, Vienna, Austria. https://epub.wu.ac.at/3558/

Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., & Mohanty, S. (2019). Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. *International Journal of Disaster Risk Reduction*, 37, 101176. https://doi.org/10.1016/j.ijdrr.2019.101176.

Řehůřek, R., & Sojka, P. (2011). *Gensim—statistical semantics in Python*. https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf.

Rossetti, M., Stella, F., & Zanker, M. (2016). Analyzing user reviews in tourism with topic models. *Information Technology & Tourism*, 16(1), 5-21. https://doi.org/10.1007/s40558-015-0035-y.

Sha, H., Hasan, M. A., Mohler, G., & Brantingham, P. J. (2020). Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives. *arXiv*. https://arxiv.org/abs/2004.11692.

Sharma, C., & Bedi, P. (2018, October 10-12). Mitigating popularity bias in Twitter- recommending novel hashtags using pooled tweets. In S. K. Niranjan (Ed.), *Proceedings of the 3rd International Conference on Contemporary Computing and Informatics (iC^{3}I - 2018)* (pp. 166-171). IEEE. https://doi.org/10.1109/IC3I44769.2018.9007248

Sharma, K., Seo, S., Meng, C., Rambhatla, S., Dua, A., & Liu, Y. (2020). COVID-19 on social media: Analyzing misinformation in Twitter conversations. *arXiv*. https://arxiv.org/abs/2003.12309.

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E., & Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv*. https://arxiv.org/abs/2003.13907.

Smith, A., Lee, T. Y., Poursabzi-Sangdeh, F., Boyd-Graber, J., Elmqvist, N., & Findlater, L. (2017). Evaluating visual representations for topic understanding and their effects on manually generated topic labels. *Transactions of the Association for Computational Linguistics*, 5, 1-16. https://doi.org/10.1162/tacl_a_00042.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012, July 12-14). Exploring topic coherence over many models and many topics. In J. Tsujii (Ed.), *EMNLP-CoNLL '12: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 952-961). Association for Computational Linguistics.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010, February 3-6). TwitterRank: Finding topic-sensitive influential Twitterers. In B. D. Davison & T. Suel (Eds.), *WSDM'10: Third ACM International Conference on Web Search and Data Mining* (pp. 261-270). ACM. https://doi.org/10.1145/1718487.1718520.

Wicke, P., & Bolognesi, M. M. (2020). Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *arXiv*. https://arxiv.org/abs/2004.06986.

World Health Organization (WHO). (2020). *Coronavirus disease (COVID-19) situation report – 142*. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200610-covid-19-sitrep-142.pdf?sfvrsn=180898cd_6.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, 59(11), 56-65. https://doi.org/10.1145/2934664.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an

appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16(Supplement 13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011, April 18-21). Comparing Twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.), *33rd European Conference on Information Retrieval Research, ECIR 2011* (pp. 338-349). Springer.

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

RESEARCH PAPER
J Inf Sci Theory Pract 9(1): 54-67, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.1.4

ISRI - Information Systems Research Constructs and Indicators: A Web Tool for Information Systems Researchers

João Varajão*
University of Minho, Centro ALGORITMI, Guimarães, Portugal
E-mail: varajao@dsi.uminho.pt

António Trigo
Polytechnic Institute of Coimbra, ISCAC, Coimbra Business School, Coimbra, Portugal
University of Minho, Centro ALGORITMI, Guimarães, Portugal
E-mail: antonio.trigo@gmail.com

Tiago Silva
University of Minho, MIEGSI,
Guimarães, Portugal
E-mail: tiagomachadosilva94@gmail.com

ABSTRACT

This paper presents the ISRI (Information Systems Research Indicators) Web tool, publicly and freely available at *isri.sciencesphere.org*. Targeting Information Systems (IS) researchers, it compiles and organizes IS adoption and use theories/models, constructs, and indicators (measuring variables) available in the scientific literature. Aiming to support the IS theory development process, the purpose of ISRI is to gather and systematize information on research indicators to help researchers and practitioners' work. The tool currently covers eleven theories/models: DeLone and McLean's IS Success Model (D&M ISS); Diffusion of Innovations Theory (DOI); Motivational Model (MM); Social Cognitive Theory (SCT); Task-Technology Fit (TTF); Technology Acceptance Model (TAM); Technology-Organization-Environment Framework (TOE); Theory of Planned Behavior (TPB); Decomposed Theory of Planned Behavior (DTPB); Theory of Reasoned Action (TRA); and Unified Theory of Acceptance and Use of Technology (UTAUT). It also includes currently over 400 constructs, nearly 2,500 indicators, and about 60 application contexts related to the models. For the creation of the tool's database, nearly 580 references were used.

Keywords: information systems adoption, information systems success, information systems theories, information systems models, information systems measuring variables, information systems scales

Received: September 17, 2020 **Accepted:** February 16, 2021

***Corresponding Author:** João Varajão
https://orcid.org/0000-0002-4303-3908
E-mail: varajao@dsi.uminho.pt

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far 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.

© João Varajão, António Trigo, Tiago Silva, 2021

1. INTRODUCTION

Information Technology (IT) and Information Systems (IS) play a central role in organizations, making it almost impossible to find an organization that does not use them. In fact, it is well accepted and somewhat consensual that IT/IS are fundamental for the development of virtually any human organization (Varajão, 2018). For instance, digitalization is one of the major trends changing business and society (Kääriäinen et al., 2020). Thus, successfully introducing new IT applications into companies is critical for developing business capabilities, contributing to organizations' competitiveness, efficiency, and profits (Patnayakuni & Ruppel, 2010; Varajão & Trigo, 2016).

Technology adoption and use theories/models seek to explain users' acceptance of new technologies and their intention to use them. In the context of IT/IS adoption, several models and theories have been applied, proposed, and developed, identifying aspects that influence the use and success of IT/IS, such as: DeLone and McLean's IS Success Model (D&M ISS) (DeLone & McLean, 2003, 2016); Task-Technology Fit (TTF) (Dishaw & Strong, 1999; Goodhue & Thompson, 1995); Technology Acceptance Model (TAM) (Davis, 1986); Technology-Organization-Environment Framework (TOE) (Tornatzky & Fleischer, 1990); or the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al., 2003), to mention a few.

It is common for these theories/models to have a high level of abstraction without restricting or detailing the measuring of the constructs that make them up, so they can be applied in several contexts. As researchers test models and theories, constructs (latent variables) are defined differently depending on the object under study (or the researchers' perspectives), making general characterization difficult, and resulting in a large quantity and diversity of indicators (measuring variables) for the same constructs. Therefore, when researchers need to define indicators for their theoretical models' constructs, it takes a long time to analyze the literature and identify references that better describe them.

For example, concerning the construct "Accuracy," different authors have used similar but different indicators (measurement variables) to represent it. Wixom and Todd (2005) define it as "represent[ing] the user's perception that the information is correct" and use the indicators: "the system produces correct information"; "there are few errors in the information I obtain from the system"; and "the information provided by the system is accurate." However, Yao and Murphy (2007) define it as "refer[ring] to the degree of correctness of the system" and use the following indicators to evaluate it: "the system will improve the efficiency"; "the system will reduce the incidence of errors when transferring data"; and "the system will reduce the incidence of errors."

No tool was found to support/help researchers develop IS theories (focusing on constructs and measuring variables). Such a tool would assist researchers' work when defining constructs and respective indicators, avoiding the need for extensive literature reviews (at least at the first phases of the research projects). To help fulfil this gap, this research aims to present the design, development, and features of a new Web tool named ISRI, which stands for "Information Systems Research Indicators." This tool should be viewed as an accelerator of the research process by facilitating the identification of relevant literature and the definition of constructs and respective indicators (measurement variables).

The paper is organized as follows: The next section presents the research and development method; section three presents a brief literature review on some of the most used theories/models; section four presents the results of the research work, the ISRI Web tool, and examples for using it in practice; and finally, section five explains the implications of using ISRI for research, as well as limitations and highlights for future research.

2. METHOD

There were two main phases for developing ISRI, as shown in Fig. 1. The first one is related to a systematic lit-

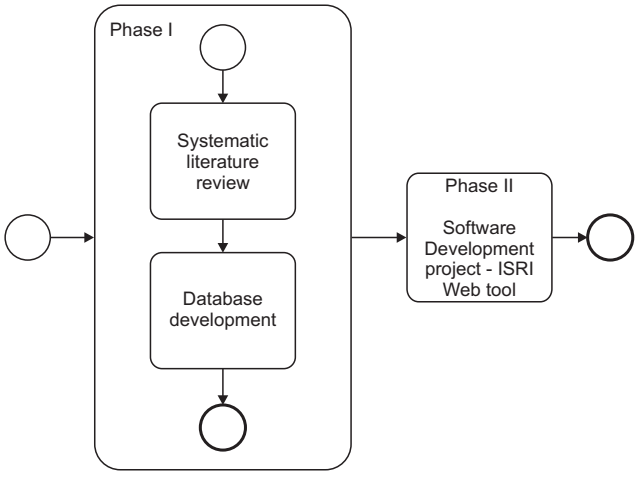


Fig. 1. Research and development method - ISRI (Information Systems Research Indicators).

erature review and the creation of the ISRI database. The second one is associated with the software development project for creating ISRI (to make available information about the theories/models gathered in the first stage).

2.1. Phase I

The first phase included a systematic literature review aiming to identify models and theories related to IT/IS adoption. Two databases were used: Elsevier's Scopus and Clarivate Analytics' (formerly Thomson-Reuters) Web of Science (WoS). From this first review, 103 theories/models were identified. A second search was then performed, focused on those models and theories, to determine the studies that used each theory/model. The selected articles were published in renowned journals in the IS field, such as *European Journal of Information Systems, Information Systems Journal, Information Systems Research, Journal of the Association for Information Systems, Journal of Information Technology, Journal of Management Information Systems, Journal of Strategic Information Systems*, and *MIS Quarterly*. Articles from conferences such as the Americas Conference on Information Systems (AMCIS), European Conference on Information Systems (ECIS), International Conference on Information Systems (ICIS), and Pacific Asia Conference on Information Systems (PACIS) have also been considered. With the information gathered, aiming to select the final theories/models for further study, it was decided to consider only those that presented a unique number of results higher than five. It enabled focusing the research on a total of eleven selected theories/models presented next.

The references were selected based on an initial analysis of the title and the abstract of the articles. Whenever the title and abstract were not sufficiently clear, general reading of the article was carried out to check if it contained the requested information. In the process of selection of articles, the following criteria were considered: focus on IT/IS adoption; use of the theories/models selected; definition of the constructs that make up the theories/models; and description of the measurement variables (indicators) of the constructs. Only articles related to the adoption, implementation, use, and acceptance of IT/IS were considered. Overall, articles that used the selected theories/models or used a theory/model resulting from the combination of at least one of the theories/models under study were considered. Articles or books where the theories/models were initially proposed were also taken into account; however, it was in a few cases impossible to access the original document.

In total, about 580 references were selected for collecting the measuring variables (indicators). The indicators identified in the articles were compiled, resulting in the database that can be consulted in ISRI.

2.2. Phase II

ISRI is being designed and developed at the University of Minho by a small team of researchers with extensive experience in IS development and project management, both in academic and industry environments. The development of the Web tool followed the typical steps of software development and related activities, namely: the specification of the software, for defining the expected tool functionalities; the design and implementation of the software, to comply with the specification; validation of the software, to test and verify that the software performs as expected and meets the expectations of the users; and the maintenance/evolution of the software, to support future needs (Sommerville, 2007).

Fig. 2 presents the functional (A) and domain models (B) defined for the project. As shown in the functional diagram (see Fig. 2A), ISRI allows users to search the theories/models, constructs, contexts, and references that support the information in the database. ISRI enables, after selecting a theory/model, viewing detailed information on the associated theory/model, such as the description, the graphical representation (diagram), the constructs, the contexts, and the related construct indicators. If the user needs to view information related to a construct or context, s/he can select it and is then presented detailed information. Regarding the domain model (see Fig. 2B), it has five classes. The main class is "TheoryModel," which represents each model/theory and, in this application context, it is understood as an aggregation of "Constructs" (latent variables). In turn, the "Construct" class is an aggregation of "Indicators" (i.e., measuring variables). The class "Indicator" is associated with the class "Reference," since the same indicator can be found in several references, and the same reference typically identifies several indicators. Finally, the "Context" class regards the contexts (e.g., "employee portal," "ERP," etc.) where each "TheoryModel" (theories/models) has been applied. ISRI has been developed using the PHP programming language, HTML 5, CSS, JavaScript, and MySQL database server.

3. THEORIES/MODELS FOR ADOPTION AND USE

The theories/models selected in the literature review are presented next in alphabetical order, as they appear

 https://doi.org/10.1633/JISTaP.2021.9.1.4

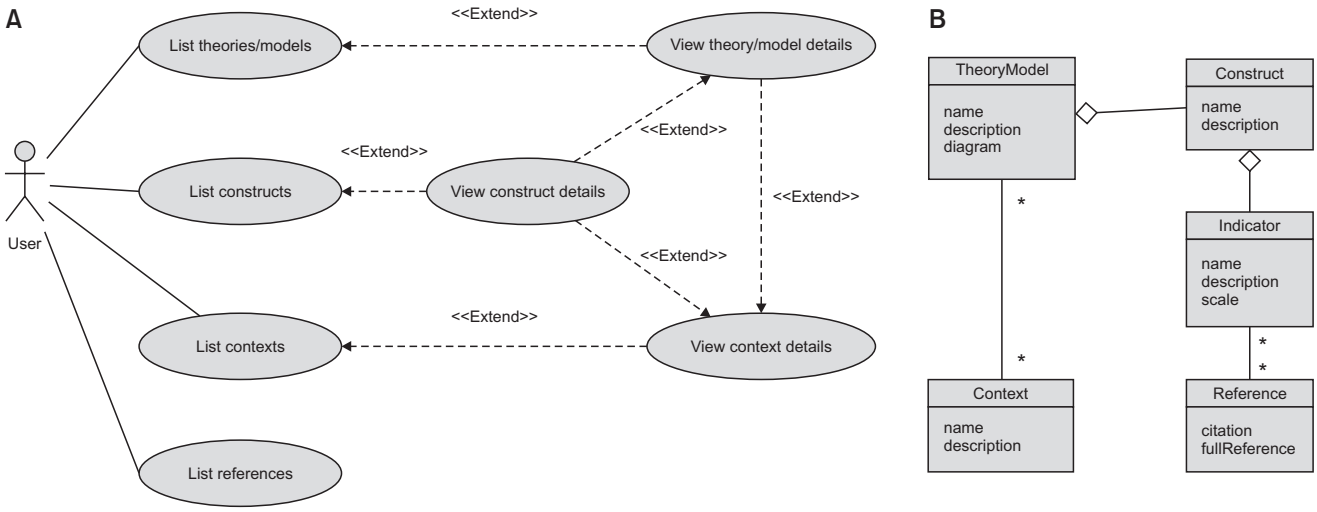


Fig. 2. Functional (A) and domain (B) models.

in the ISRI: Decomposed Theory of Planned Behavior (DTPB) (Taylor & Todd, 1995b); D&M ISS (DeLone & McLean, 2003, 2016); Diffusion of Innovations Theory (DOI) (Rogers, 1962, 2003); Motivational Model (MM) (Davis et al., 1992); Social Cognitive Theory (SCT) (Bandura, 1986); TTF (Dishaw & Strong, 1999; Goodhue & Thompson, 1995); TAM (Davis, 1986); TOE (Tornatzky & Fleischer, 1990); Theory of Planned Behavior (TPB) (Ajzen, 1985, 1991); Theory of Reasoned Action (TRA) (Fishbein & Ajzen, 1975); and the UTAUT (Venkatesh et al., 2003). For each theory/model, a brief description is presented. A more detailed description of each theory/model, including constructs, indicators, contexts, and diagram, is available in ISRI and respective sources.

3.1. Decomposed Theory of Planned Behavior

The DTPB, proposed by Taylor and Todd (1995b), consists of a decomposition of TPB (Ajzen, 1991). This theory was proposed to represent better the relationships between belief structures and antecedents of intention (Taylor & Todd, 1995b). The decomposition made to TPB (Ajzen, 1991) focused on the Attitude, Subjective Norm, and Perceived Behavioral Control constructs. According to Taylor and Todd (1995b), the decomposition of beliefs has advantages, such as clarifying and improving the understanding of existing relationships, providing beliefs that can be used in different contexts. The model becomes administratively relevant by focusing on specific beliefs, pointing to specific factors that may influence adoption and use. These factors can be worked through system design and implementation strategies.

3.2. DeLone and McLean IS Success Model

The D&M ISS, proposed by DeLone and McLean (2003), is based on theoretical and empirical research in IS, conducted by several researchers between 1981 and 1990 (DeLone & McLean, 2016). According to DeLone and McLean (1992), the D&M ISS was proposed as a framework and model to conceptualize and operationalize IS success. In 2003, the authors proposed an update to the model, becoming known as the Updated D&M ISS. This update emerged after the authors reviewed the various contributions on the success of IS over the previous decade, focusing essentially on research efforts that applied, validated, challenged, and proposed improvements to the original model (DeLone & McLean, 2003).

3.3. Diffusion of Innovations Theory

The DOI was founded in sociology and popularized by Rogers (1962, 1983, 1995, 2003). Rogers synthesized many studies in diffusion from fields that influenced the theory, including anthropology, early sociology, rural sociology, education, industrial sociology, and medical sociology. With the synthesis accomplished, Rogers produced a theory of the adoption of innovations between individuals and organizations (Rogers, 1962).

3.4. Motivational Model

The MM was presented by Davis et al. (1992) in the context of the use of technology. The MM theorizes that extrinsic and intrinsic motivation are the main drivers of an individual's intention to perform a given behavior, and the intention to perform a behavior is a construct that is

closely linked to the actual behavior (Venkatesh & Speier, 1999). Davis et al. (1992) argue that people's intentions to use computers in the workplace are influenced mainly by their perception of how useful computers are to improve their work performance, and secondarily by the degree of pleasure they have in using the computers themselves.

3.5. Social Cognitive Theory

The SCT, founded in psychology, is a theory that provides a framework for understanding, predicting, and changing human behavior (Bandura, 1986). This theory was proposed by Bandura (1986), and it was first named Social Learning Theory. The SCT states that human functioning is the product of reciprocal interactions of intrapersonal, behavioral, and environmental determinants (Bandura, 1986, 2012). Subsequently, Compeau and Higgins (1995) applied and extended SCT to the context of computer use. However, the nature of the model and the underlying theory allow it to be extended to the acceptance and use of IT in general (Venkatesh et al., 2003). This model proposes the use of several constructs, such as self-efficacy, anxiety, outcome expectations (usefulness), and affects (attitude toward using technology) (Rana & Dwivedi, 2015). According to Compeau and Higgins (1995), encouragement by others, the others' use, and organizational support in computer use, influence computer self-efficacy and outcome expectations. Computer self-efficacy influences usage directly as well as indirectly through outcome expectations, affect, and anxiety. Outcome expectations influence usage directly as well as indirectly through affect. Finally, affect and anxiety are influential in usage.

3.6. Task-Technology Fit

The TTF, proposed by Goodhue and Thompson (1995), states that user use and attitude towards technology impact individual performance. According to Dishaw and Strong (1999), the TFF model attempts to solve one of the TAM's main weaknesses concerning the understanding of IT use. The weakness in question is the lack of focus on the task, since IT is the tool that the user uses to carry out their tasks. This weakness contributed to the variety in the results of IT evaluations (Goodhue & Thompson, 1995). Although TAM's concept of utility is focused on tasks, the addition of more task features could provide a better model for IT use (D'Ambra et al., 2013). According to Dishaw and Strong (1999), the TTF focuses on the correspondence between the user's task needs and the available IT functionalities. Goodhue and Thompson (1995) proposed the Technology-to-Performance Chain, where IT characteristics, tasks, and individual users explain the use of the information system and individual performance.

3.7. Technology Acceptance Model

The TAM was initially presented by Davis (1986) while conducting doctoral thesis work. This model is an adaptation of Theory TRA (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975), where the main objective is to explain the behavior of the use of IS (Davis et al., 1989). Later in 1989, with the work developed by Davis et al. (1989), the TAM was finally conceptualized, gaining visibility. The TAM intends to explain computer system users' behavior, claiming that the use of the computer system depends directly on the individual's intention in using it. TAM refers to Behavioral Intention to Use, which results from the influence of the user's Attitude Toward Using and Perceived Usefulness. As a result, Attitude Toward Using is influenced by two essential factors, Perceived Usefulness and Perceived Ease of Use. Davis (1986) consider that the Perceived Ease of Use can also affect the Perceived Usefulness of the system, while External Variables influence both. The most current version is TAM 3, which presents a complete nomological network (integrated model) of the determinants of IT adoption by individuals (Venkatesh & Bala, 2008).

3.8. Technology-Organization-Environment Framework

The TOE Framework was developed by Tornatzky and Fleischer (1990) and is consistent with the DOI (Rogers, 1983). The TOE framework proved to be useful for studying the adoption and assimilation of different IT innovation types (Oliveira & Martins, 2010). The TOE is a useful and adaptive framework to explain adoption behavior concerning three types of technological innovations, namely applied innovations for technical tasks, innovations for business administration, and innovations embedded in the core business processes of an organization (Ramdani & Kawalek, 2007; Swanson, 1994). The TOE framework identifies three aspects of an organization's context that influence adopting and implementing a technological innovation, namely technological context, organizational context, and external environmental context.

3.9. Theory of Planned Behavior

The TPB proposed by Ajzen (1985, 1991) is an extension of the TRA (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975), made necessary from the limitations of the original model in dealing with behaviors over which peo-

 https://doi.org/10.1633/JISTaP.2021.9.1.4

ple have incomplete volitional control (Ajzen, 1991). The TPB proposes that the intention of an individual to have a given behavior can be explained with precision by the attitude that the individual has to the behavior, by a set of subjective norms, and by the perception about the control of the behavior by the individual (Ajzen, 1991).

3.10. Theory of Reasoned Action

The TRA came from the social psychology area and was proposed by Fishbein and Ajzen (1975). TRA is one of the most fundamental and influential theories of human behavior (Venkatesh et al., 2003), where the objective is to investigate the relationship between attitude and behavior based on two main concepts, principles of compatibility and behavioral intention (Fishbein & Ajzen, 1975; Mishra et al., 2014). According to TRA, the individual's performance in a particular behavior is determined by his or her Behavioral Intention to perform the behavior, and that Behavioral Intention is determined jointly by the person's Attitude Towards Behavior and Subjective Norm on the behavior in question. TRA is not especially related to the use of IT. However, the use of IT is a special case of general behavior. Thus, it is easy to understand why TRA underlies theories such as TAM or UTAUT.

3.11. Unified Theory of Acceptance and Use of Technology

The UTAUT, proposed by Venkatesh et al. (2003), was formulated through a combination of eight models of IT acceptance research, including TRA, TAM, MM, TPB, a model combining the TAM and TPB (Taylor & Todd, 1995c), Model of PC Utilization (Thompson et al., 1991), DOI, and SCT. According to Venkatesh et al. (2003), Performance Expectancy, Effort Expectancy, and Social Influence were formulated to influence Behavioral Intention in the use of technology, while Behavioral Intention and Facilitating Conditions determine Use Behavior. Individual difference variables such as Age, Gender, Experience, and Voluntariness of Use were also formulated to moderate the various relationships in UTAUT.

4. RESULTS

4.1. Overview of the User Interface

ISRI aims to support IS researchers in the process of theories/models' study, development, and testing, by organizing and presenting theories/models, constructs, and contexts, focusing on the constructs' measuring variables (indicators). In ISRI designing, care was taken to assure that it is intuitive and easy to use. In this sense, all the information in ISRI can be accessed from the initial homepage, as presented in Fig. 3.

ISRI has a navigation menu at the top that allows the user to access the perspectives by which s/he wants to search the database, namely: theories/models, constructs, contexts, or references. The main page also has a menu in the center with the main search perspectives (theories/models, constructs, and contexts). When the user clicks on one of these perspectives, s/he will see on the right side the elements related to them. By clicking the theories/models option, the user is taken to the theories/models page, being presented the list of options available, which currently has eleven theories/models. After selecting the required theory/model, the user is taken to the respective web page. This page has a menu that enables access to the main aspects of the theory/model, namely the description, the graphical diagram, the constructs, contexts of use, and the indicators by construct, as exemplified in Fig. 4.

Fig. 4A shows the description of a selected model and the respective references. In this example, the model shown is the DTPB. By clicking on the "Diagram" tab, the user has access to the diagram of the model (Fig. 4B). The

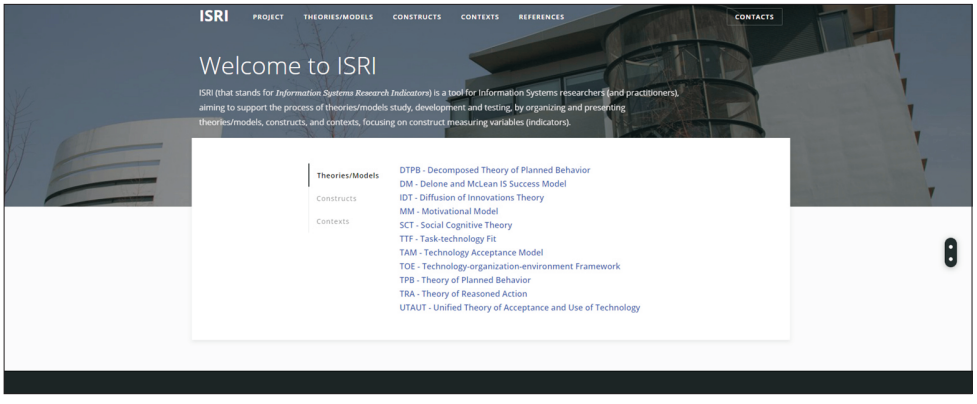


Fig. 3. ISRI (Information Systems Research Indicators) home page (available at *isri.sciencesphere.org*).

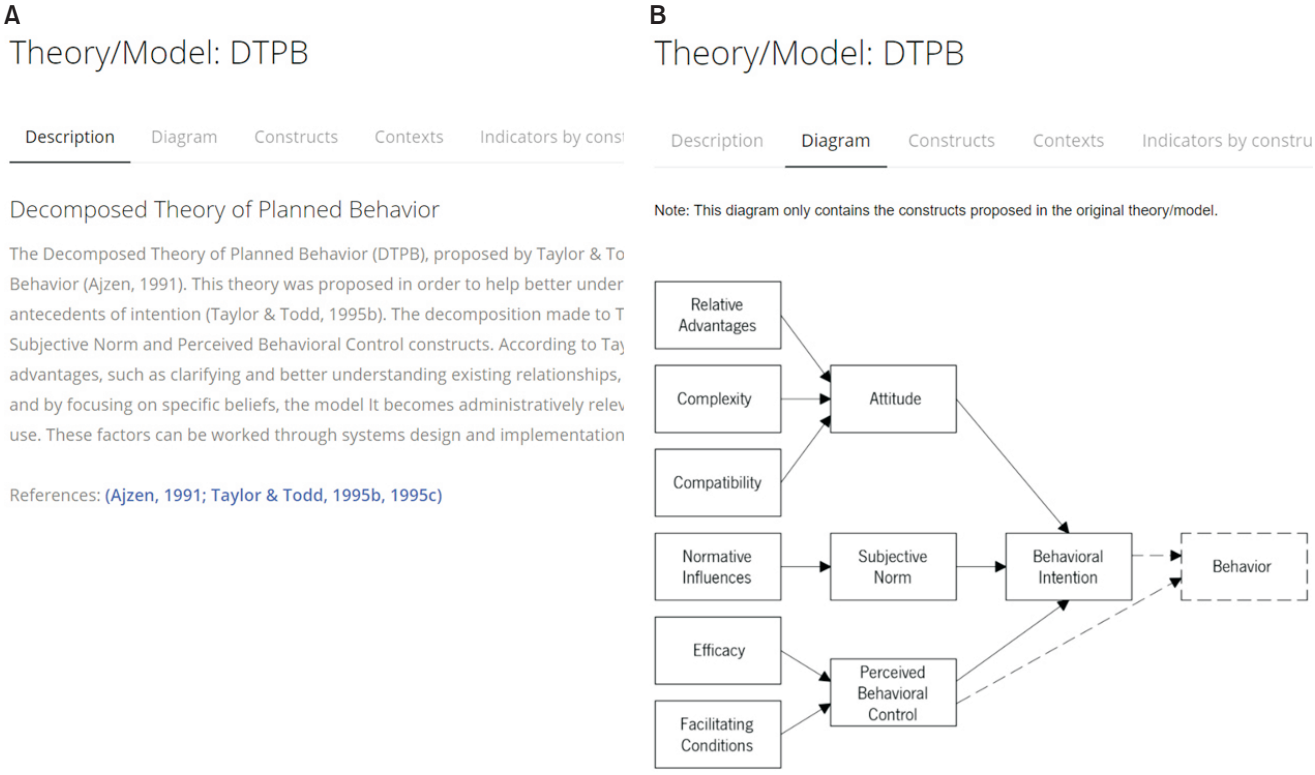


Fig. 4. Example of theory/model description (A) and corresponding diagram (B) in ISRI (Information Systems Research Indicators).

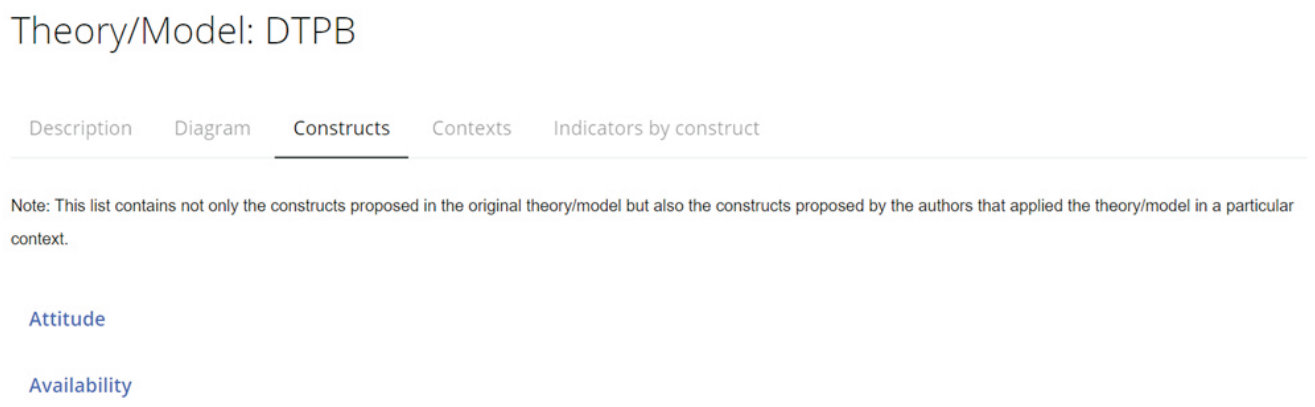


Fig. 5. Example of constructs of a selected theory/model in ISRI (Information Systems Research Indicators).

diagram provides a visual overview of the theory/model. By clicking on the "Constructs" tab, the user has access to the list of all the constructs of the selected model (Fig. 5). For instance, in the case of DTPB, the model has 30 associated constructs which are (see *isri.sciencesphere.org*): Attitude; Availability; Behavioral Intention; Behavioral Intention for Continued Use; Collaborative Norms; Compatibility; Complexity; Ease of Use; Efficacy; Facilitating Conditions; Facilitating Conditions (Resources); Facilitating Conditions (Technology); Family, Relatives, Friends, and Peer Influence; Future Obligation; Governmental Influence; Hedonic Outcomes; Normative Influences; Peer Influence; Perceived Behavioral Control; Perceived Ease of Use; Perceived Usefulness; Personal Network Exposure; Relative Advantage; Seeker Knowledge Growth; Self-Efficacy; Subjective Norm; Superior's Influence; Usage; and Utilitarian Outcomes. Please note that it includes constructs from the original model and the extended models/

 https://doi.org/10.1633/JISTaP.2021.9.1.4

references that applied to the original model and added new constructs.

It is possible for each construct to find several definitions, as presented by different authors, as well as the list of associated indicators. For each indicator listed for a given construct, there is shown the text related to the indicator (the original question that was asked), the scale used to measure that indicator, the context in which it was used, and the respective references. Fig. 6 shows an example of the construct "Complexity."

By clicking on the "Contexts" tab (see Fig. 4), the user has access to contexts in which the model has already been applied. This perspective in ISRI is important to identify the theories/models already used to study specific contexts. In the case of the DTPB model in Fig. 7, it is noticeable that it has already been used in the following contexts: Computing Resource Center; Electronic Knowledge Repository; Internet TV; and VCR Plus and TM. A significant advantage of this perspective is the possibility of analyzing the similarities and differences in the definitions of the constructs when the theories/models have been applied in different contexts. On the other hand, it is also possible to verify differences, sometimes significant, even when theories/models are applied to the same contexts.

Returning to the main menu (Fig. 3) it is possible, as previously mentioned, to search ISRI for the other two perspectives: constructs and contexts. According to the selected perspective, a list of all available items will appear (by clicking on "Constructs" or "Contexts"). For instance, the user may want to check if a theory/model has already been applied in a given context (e.g., e-Government). There is also the last tab, the one for references, where it is possible to list, in full, all the used references.

4.2. Demonstration

For demonstration purposes, we first searched the Scopus search engine, looking for papers that had in the title (or abstract) the name of any of the models available on ISRI (e.g., "Technology Acceptance Model"). The idea was to find good examples for using ISRI (i.e., where ISRI could have contributed to the initial development of the presented theory/model). From the results obtained, two papers were selected since they present the theory/model, the constructs and indicators used, and were applied in two different contexts. The first selected paper, entitled "Student acceptance of virtual laboratory and practical work: An extension of the technology acceptance model" (Estriegana et al., 2019), uses TAM in the context of e-learning. The second one, entitled "Factors influencing the adoption of mHealth services in a developing country: A patient-centric study" (Alam et al., 2020), uses the UTAUT

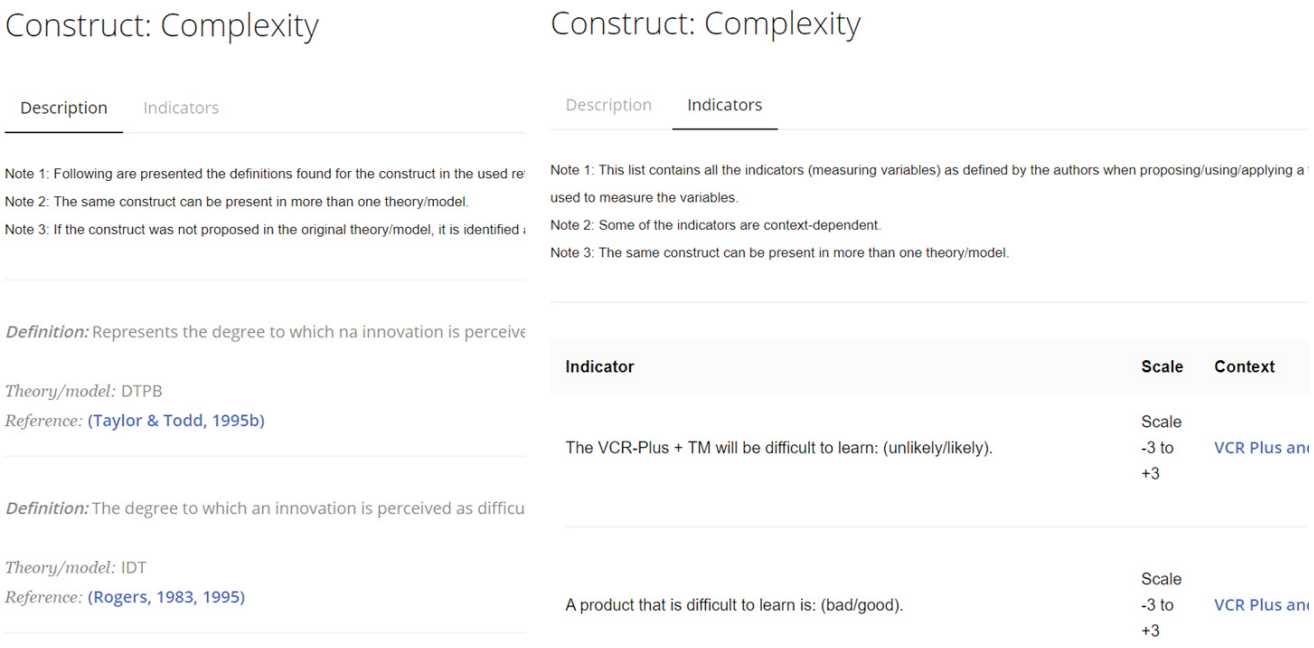


Fig. 6. Example of a selected construct description and associated indicators (measuring variables) in ISRI (Information Systems Research Indicators).

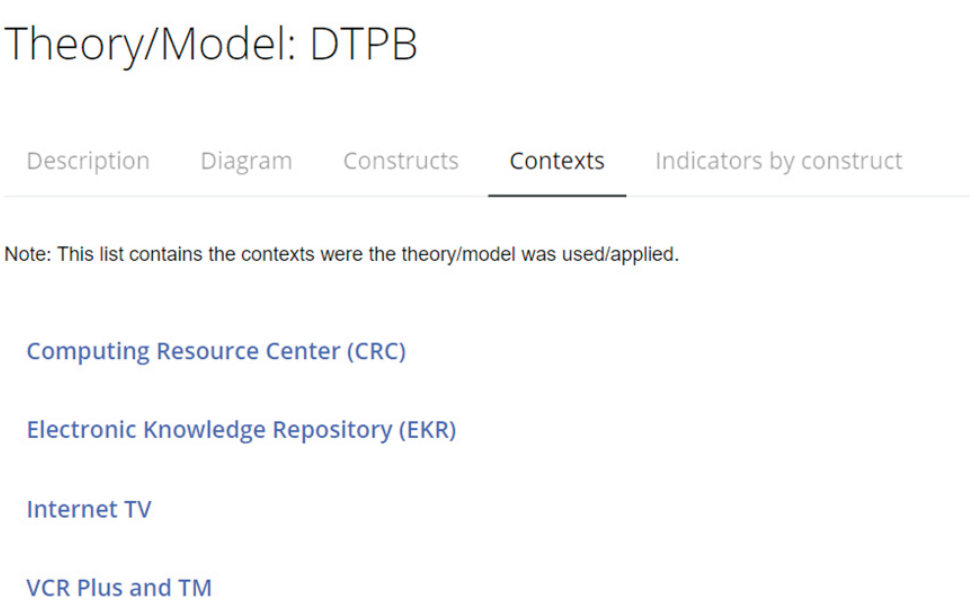


Fig. 7. Contexts of application of a selected theory/model.

model in the context of mobile computing. Please note that these two papers were randomly selected for illustrative purposes and do not relate to this work or ISRI.

4.2.1. An example regarding the paper "Student acceptance of virtual laboratory and practical work: An extension of the technology acceptance model"

The paper of Estriegana et al. (2019) presents the results of an empirical survey study whose purpose was to examine students' acceptance of technology and the process of adopting an online learning environment incorporating Web-based resources, such as virtual laboratories, interactive activities, and educational videos, and a game-based learning methodology using TAM.

Suppose the authors had chosen to conduct their study with the help of ISRI. In that case, they could have started their research by selecting the context, which is the IT adoption associated with e-learning. After selecting the context, the researchers would have access to the available constructs, which in this case are the constructs associated with the TAM model (as can be found at *isri.sciencesphere.org*): Attitude Towards Use; Computer Self-efficacy; Intention to Use; Perceived Ease of Use; Perceived Enjoyment; Perceived Usefulness; Actual Usage; Cognitive Absorption; Content Quality; External Influence; Internet Self-efficacy; Interpersonal Influence; Learning Goal Orientation; Network Externality; Perceived Performance; System Functionality; System Interactivity; and System Response. In ISRI, for each construct, there are available indicators reported in the literature to measure it. For example, the indicators associated with the first construct, Attitude Towards Use, are presented in Fig. 8.

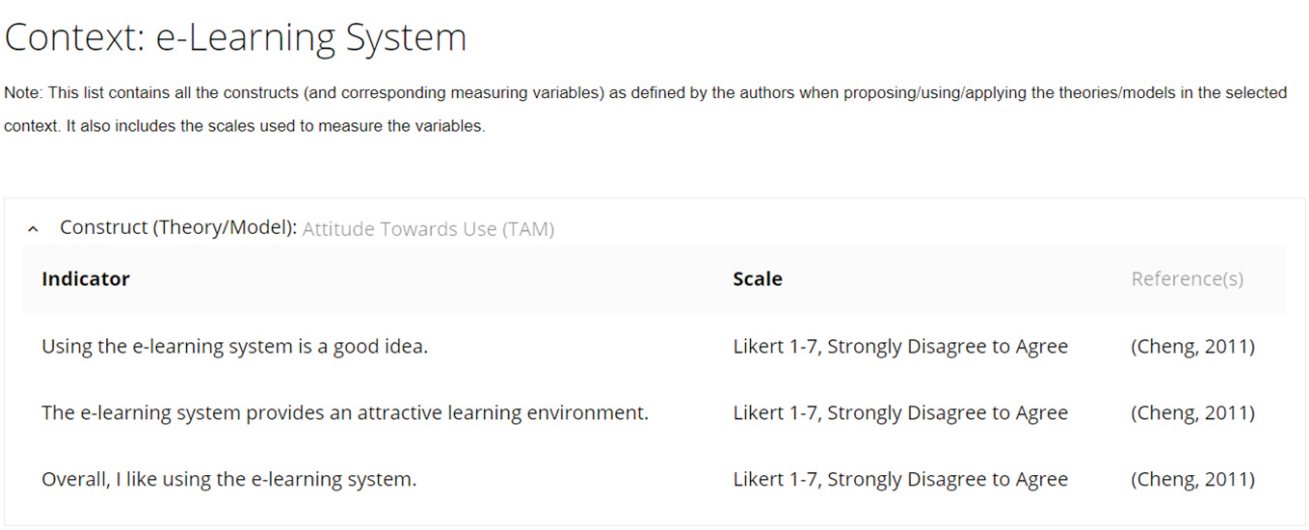
Following this, considering the model of their study (TAM), the authors could have identified the indicators of all the constructs available in ISRI: Attitude Toward Using; Behavioral Intention to Use; Efficiency; Perceived Ease of Use; Perceived Usefulness; Playfulness; and a new construct, Perceived Satisfaction. Although the work developed by the authors may involve other references than those available at ISRI, the use of ISRI is useful in the identification of an initial version of the model to fasten the research process.

4.2.2. An example regarding the paper "Factors influencing the adoption of mHealth services in a developing country: A patient-centric study"

The objective of Alam et al. (2020) was to study the adoption of mHealth services in a developing country using the UTAUT model. In this case, mHealth services are defined as "medical and public health practices supported by mobile devices, such as mobile phones, that are used to make decisions and provide emergency care and, among other benefits, that help the elderly to manage their daily activities associated with independent living."

As in the previous example, if the authors had started their research by searching ISRI by context, in this case the use of mobile devices in health, they would have

 https://doi.org/10.1633/JISTaP.2021.9.1.4



Context: e-Learning System

Note: This list contains all the constructs (and corresponding measuring variables) as defined by the authors when proposing/using/applying the theories/models in the selected context. It also includes the scales used to measure the variables.

Construct (Theory/Model): Attitude Towards Use (TAM)

Indicator	Scale	Reference(s)
Using the e-learning system is a good idea.	Likert 1-7, Strongly Disagree to Agree	(Cheng, 2011)
The e-learning system provides an attractive learning environment.	Likert 1-7, Strongly Disagree to Agree	(Cheng, 2011)
Overall, I like using the e-learning system.	Likert 1-7, Strongly Disagree to Agree	(Cheng, 2011)

Fig. 8. Indicators (measuring variables) available in ISRI (Information Systems Research Indicators), related to the construct *Attitude Towards Use* in the *e-Learning System* context.

Table 1. Comparison of indicators (measuring variables) for the *Behavioral Intention* construct

Construct: *Behavioral Intention*			
ISRI Web tool		Study of Alam et al. (2020)	
Indicator	References	Indicator	References
I intend to continue using mobile Internet in the future	(Venkatesh et al., 2012)	I intend to continue using mHealth in the future	(Taylor & Todd, 1995a; Venkatesh et al., 2012)
I will always try to use mobile Internet in my daily life	(Venkatesh et al., 2012)	I will always try to use mHealth in my daily life	(Taylor & Todd, 1995a; Venkatesh et al., 2012)
I plan to continue to use mobile Internet frequently	(Venkatesh et al., 2012)	I plan to continue to use mHealth services	(Taylor & Todd, 1995a; Venkatesh et al., 2012)

ISRI, Information Systems Research Indicators.

found that a possible model for their study is the UTAUT model. Although there is no "mHealth" context in ISRI, there are four contexts associated with mobile devices/services (i.e., Mobile Business [m-business], Mobile Devices, Mobile Internet, and Mobile Technologies). There are 55 constructs related to the UTAUT model in ISRI, and researchers can view details on the indicators for each construct. The work done by the authors identifies the following constructs (Alam et al., 2020): Performance Expectancy; Effort Expectancy; Social Influence; Facilitating Condition; Perceived Reliability; Price Value; Gender; Behavioral Intention; and Actual Use Behavior, which are all available in ISRI. When comparing the indicators that the authors of this second paper have chosen to measure the constructs (e.g., Behavioral Intention construct) to the ones returned by ISRI, it turns out that they are similar, as it is possible to see in Table 1.

Since all the authors' constructs/indicators to perform their work are available on ISRI (with the appropriate adaptations), ISRI could also have been very useful in this context.

One can argue that both presented examples were successfully developed without the need for ISRI. This is a fact; however, the purpose of ISRI is not to replace the researchers or the literature review process. On the contrary, it aims to support that work by making the process easier and more straightforward.

4.2.3. Hypothetical example

This section presents an example of the use of ISRI based on a hypothetical example. The aim is to describe a case of using ISRI in the first stages of the research work.

Suppose that a new theoretical model like the one presented in Fig. 9 must be developed.

The model has three constructs (Anxiety, Hedonic motivation, and Behavioral Expectation). ISRI can help develop this model in several aspects: 1) Definition of constructs; 2) Identification of measuring variables/indicators; 3) Identification of the relevant literature.

Regarding the definition of constructs, by selecting the "Constructs" perspective, the user obtains the list of constructs and respective definitions, as it is possible to verify in Fig. 10.

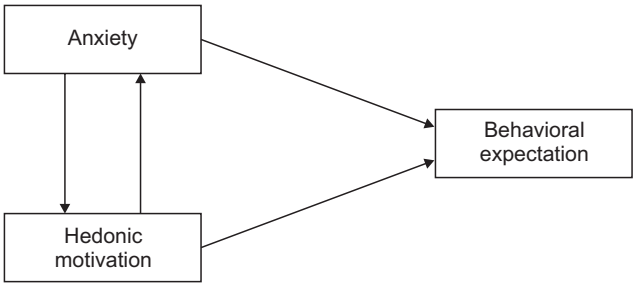


Fig. 9. Hypothetical research model.

Regarding the identification of measuring variables (indicators), by selecting each construct, the user can go to the "Indicators" tab and obtain the available indicators in ISRI, which are: for Anxiety: "*I feel apprehensive about using [a] computer*"; "*It scares me to think that I could cause the computer to destroy a large amount of information by hitting the wrong key*"; "*I hesitate to use a computer for fear of making mistakes I cannot correct*"; "*Computers are somewhat intimidating to me*"; "*I feel apprehensive about using the system*"; "*It scares me to think that I could lose a lot of information using the system by hitting the wrong key*"; "*I hesitate to use the system for fear of making mistakes I cannot correct*"; and "*The system is somewhat intimidating to me*"; for Hedonic Motivation: "*Using mobile Internet is fun*"; "*Using mobile Internet is enjoyable*"; and "*Using mobile Internet is very entertaining*"; and for Behavioral Expectation: "*I expect to use the system in the next -n- months*"; "*I will use the system in the next -n- months*"; and "*I am likely to use the system in the next -n- months*."

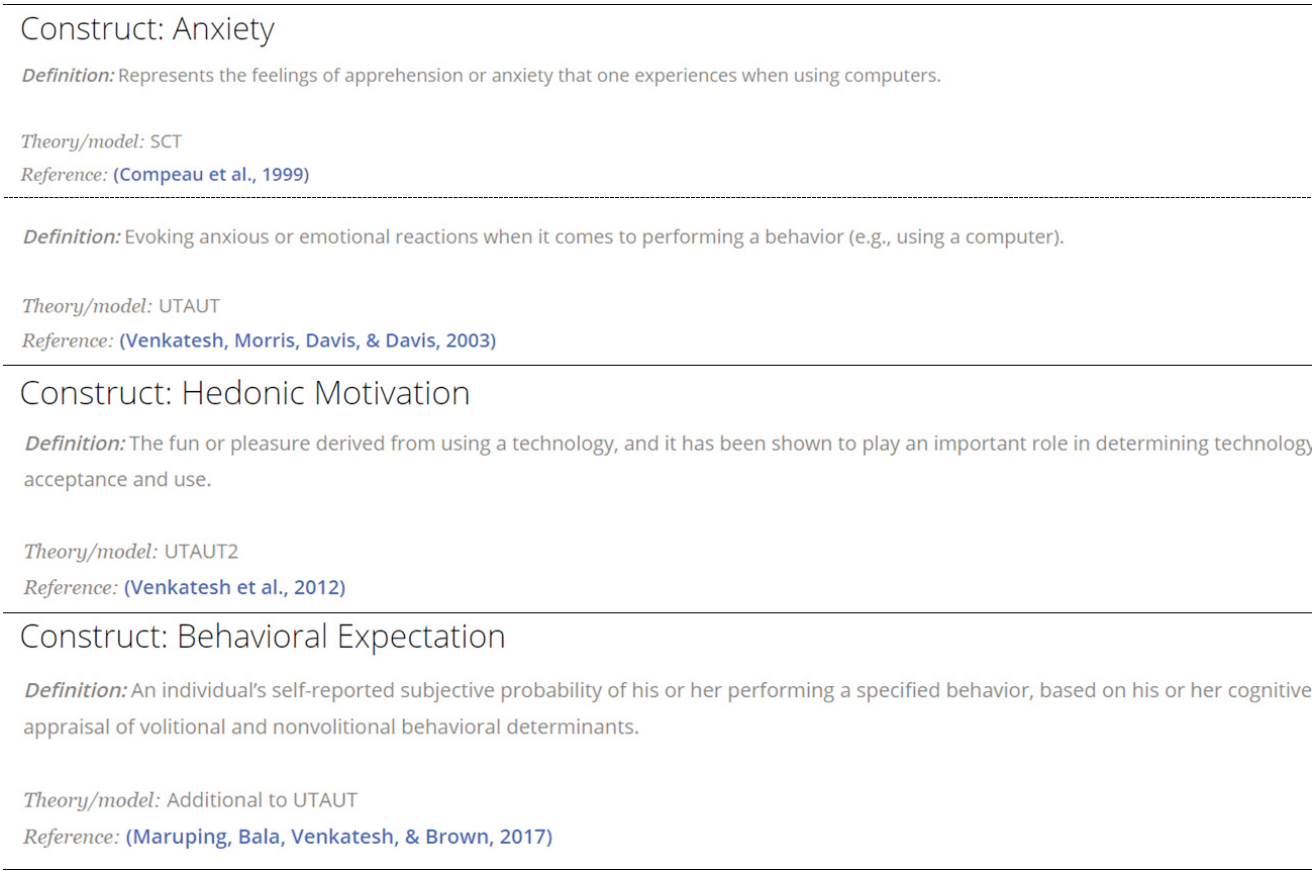


Fig. 10. Definitions available in ISRI (Information Systems Research Indicators) for the selected constructs.

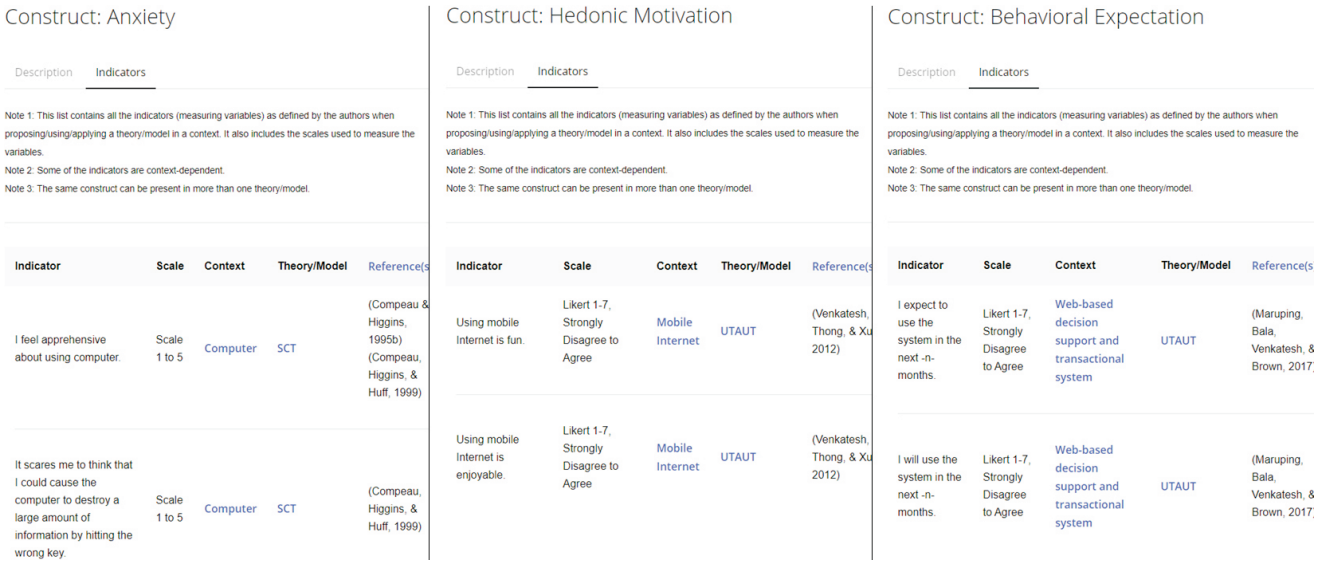


Fig. 11. Sample of available indicators (measuring variables) in ISRI (Information Systems Research Indicators) for the constructs *Anxiety*, *Hedonic Motivation and Behavioral Expectation* (see *isri.sciencesphere.org*).

Fig. 11 provides an extract of the indicators identified above, organized by construct. As shown in the figure, for each indicator is available: the description of the indicator; the measuring scale; the context and the theory/model in which it was used; and the respective references.

Finally, the user can access the full references related to the constructs/indicators by clicking on them.

5. DISCUSSION AND FUTURE DEVELOPMENTS

The work presented in this paper addresses a common challenge in research work, which is related to the time consumed in the early stages of the literature review process for defining and developing theories/models (in this case, specifically associated with IS).

There are two main contributions from ISRI. The first one is the ISRI database of theories/models, which resulted from a systematic literature review. Eleven theories/models, over 400 constructs, nearly 2,500 indicators, and about 60 contexts of the application were identified, compiled from 580 references. The second contribution is the ISRI Web tool itself, which quickly and intuitively provides the information stored in the database, allowing researchers to access it from different perspectives: theories/models, constructs, and contexts.

Another contribution of this work is related to providing examples for using ISRI in practice, presented in section four. The first two examples show how two recently published research papers in leading journals could have used ISRI as a source of information to develop the applied models. The third example shows how ISRI can be used to support the development of a new research model.

The main limitation of ISRI is that it is always an ongoing work: The increasing rate of publication of articles in the IT/IS adoption and use topics requires constant updating of the ISRI database. There is also space to expand the tool functionalities, including advanced search features.

Finally, we would like to stress that ISRI is not only a tool to facilitate research. It is also a tool for support learning since it provides students and young researchers with a set of the most applied theories/models and respective sources in IT/IS adoption and use.

As future work, we propose developing new functionalities to be added to ISRI; for instance, regarding the measurement variables (indicators) collection and presentation. It is also suggested to carry out case studies to identify opportunities for further development.

CONFLICTS OF INTEREST

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

REFERENCES

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl, & J. Beckmann (Eds.), *Action control: From cognition to behavior* (pp. 11-39). Springer.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. https://doi.org/10.1016/0749-5978(91)90020-T.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Prentice-Hall.

Alam, M. Z., Hoque, M. R., Hu, W., & Barua, Z. (2020). Factors influencing the adoption of mHealth services in a developing country: A patient-centric study. *International Journal of Information Management*, 50, 128-143. https://doi.org/10.1016/j.ijinfomgt.2019.04.016.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bandura, A. (2012). Social cognitive theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 349-374). Sage. https://doi.org/10.4135/9781446249215.n18.

Compeau, D., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 19(2), 189-211. https://doi.org/10.2307/249688.

D'Ambra, J., Wilson, C. S., & Akter, S. (2013). Application of the task-technology fit model to structure and evaluate the adoption of E-books by academics. *Journal of the American Society for Information Science and Technology*, 64(1), 48-64. https://doi.org/10.1002/asi.22757.

Davis, F. D. Jr. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* (PhD thesis). Massachusetts Institute of Technology, Cambridge, MA, USA. http://hdl.handle.net/1721.1/15192

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982-1003. https://www.jstor.org/stable/2632151?seq=1.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology*, 22(14), 1111-1132. https://doi.org/10.1111/j.1559-1816.1992.tb00945.x.

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, 3(1), 60-95. https://doi.org/10.1287/isre.3.1.60.

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*, 19(4), 9-30. https://doi.org/10.1080/07421222.2003.11045748.

DeLone, W. H., & McLean, E. R. (2016). Information systems success measurement. *Foundations and Trends® in Information Systems*, 2(1), 1-116. https://doi.org/10.1561/2900000005.

Dishaw, M. T., & Strong, D. M. (1999). Extending the technology acceptance model with task–technology fit constructs. *Information & Management*, 36(1), 9-21. https://doi.org/10.1016/S0378-7206(98)00101-3.

Estriegana, R., Medina-Merodio, J.-A., & Barchino, R. (2019). Student acceptance of virtual laboratory and practical work: An extension of the technology acceptance model. *Computers & Education*, 135, 1-14. https://doi.org/10.1016/j.compedu.2019.02.010.

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention and behavior: *An introduction to theory and research*. Addison-Wesley.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, 19(2), 213-236. https://doi.org/10.2307/249689.

Kääriäinen, J., Pussinen, P., Saari, L., Kuusisto, O., Saarela, M., & Hänninen, K. (2020). Applying the positioning phase of the digital transformation model in practice for SMEs: Toward systematic development of digitalization. *International Journal of Information Systems and Project Management*, 8(4), 24-43. https://doi.org/10.12821/ijispm080402.

Mishra, D., Akman, I., & Mishra, A. (2014). Theory of reasoned action application for green information technology acceptance. *Computers in Human Behavior*, 36, 29-40. https://doi.org/10.1016/j.chb.2014.03.030.

Oliveira, T., & Martins, M. F. (2010, September 9-10). Information technology adoption models at firm level: Review of literature. In M. de Castro Neto (Ed.), *Proceedings of the 4th European Conference on Information Management and Evaluation* (*ECISME 2010*) (pp. 312-323). Academic Publishing.

Patnayakuni, R., & Ruppel, C. P. (2010). A socio-technical approach to improving the systems development process. *Information Systems Frontiers*, 12(2), 219-234. https://doi.org/10.1007/s10796-008-9093-4.

Ramdani, B., & Kawalek, P. (2007, June 14-16). SME adoption of enterprise systems in the Northwest of England. In T. McMaster, D. Wastell, E. Ferneley, & J. I. DeGross (Eds.), *Organizational dynamics of technology-based innovation: Diversifying the research agenda. TDIT 2007* (pp. 409-429). Springer. https://doi.org/10.1007/978-0-387-72804-9_27

Rana, N. P., & Dwivedi, Y. K. (2015). Citizen's adoption of an e-government system: Validating extended social cognitive theory (SCT). *Government Information Quarterly*, 32(2), 172-181. https://doi.org/10.1016/j.giq.2015.02.002.

Rogers, E. M. (1962). *Diffusion of innovations*. Free Press of Glencoe.

Rogers, E. M. (1983). *Diffusion of innovations* (*3rd ed*). Free

Press.

Rogers, E. M. (1995). *Diffusion of innovations* (*4th ed*). Free Press.

Rogers, E. M. (2003). *Diffusion of innovations* (*5th ed*). Free Press.

Sommerville, I. (2007). *Software engineering* (*8th ed*). Addison-Wesley.

Swanson, E. B. (1994). Information systems innovation among organizations. *Management Science*, 40(9), 1069-1092. https://www.jstor.org/stable/2661474.

Taylor, S., & Todd, P. (1995a). Assessing IT usage: The role of prior experience. *MIS Quarterly*, 19(4), 561-570. https://doi.org/10.2307/249633.

Taylor, S., & Todd, P. (1995b). Decomposition and crossover effects in the theory of planned behavior: A study of consumer adoption intentions. *International Journal of Research in Marketing*, 12(2), 137-155. https://doi.org/10.1016/0167-8116(94)00019-K.

Taylor, S., & Todd, P. (1995c). Understanding information technology usage: A test of competing models. *Information Systems Research*, 6(2), 144-176. https://www.jstor.org/stable/23011007.

Thompson, R. L., Higgins, C. A., & Howell, J. M. (1991). Personal computing: Toward a conceptual model of utilization. *MIS Quarterly*, 15(1), 125-143. https://doi.org/10.2307/249443.

Tornatzky, L. G., & Fleischer, M. (1990). *The processes of technological innovation*. Lexington Books.

Varajão, J. (2018). The many facets of information systems (+projects) success. *International Journal of Information Systems and Project Management*, 6(4), 5-13. https://doi.org/10.12821/ijispm060401.

Varajão, J., & Trigo, A. (2016, December 11-14). Evaluation of IS project success in InfSysMakers: An exploratory case study. *37th International Conference on Information Systems* (*ICIS 2016*) (pp. 1-10). Association for Information Systems.

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, 39(2), 273-315. https://doi.org/10.1111/j.1540-5915.2008.00192.x.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478. https://doi.org/10.2307/30036540.

Venkatesh, V., & Speier, C. (1999). Computer technology training in the workplace: A longitudinal investigation of the effect of mood. *Organizational Behavior and Human Decision Processes*, 79(1), 1-28. https://doi.org/10.1006/obhd.1999.2837.

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36(1), 157-178. https://doi.org/10.2307/41410412.

Wixom, B. H., & Todd, P. A. (2005). A theoretical integration of user satisfaction and technology acceptance. *Information Systems Research*, 16(1), 85-102. https://www.jstor.org/stable/23015766.

Yao, Y., & Murphy, L. (2007). Remote electronic voting systems: An exploration of voters' perceptions and intention to use. *European Journal of Information Systems*, 16(2), 106-120. https://doi.org/10.1057/palgrave.ejis.3000672.

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

RESEARCH PAPER
J Inf Sci Theory Pract 9(1): 68-78, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.1.5

The Use of Weblogs as a Tool for Thai Political Engagement

Sutthinan Chuenchom*
Department of Information Science, Faculty of Humanities and Social Sciences, Chiang Mai Rajabhat University, Chiang Mai, Thailand
E-mail: sutthinan@cmru.ac.th

ABSTRACT

Political weblogs are as diverse as political viewpoints are. In the period of Prime Minister Yingluck Shinawatra, several political crises occurred, such as opposition to the Amnesty Act, the constitutional amendment, and the anti-government protests. Remarkably, during this time, social media were used as a platform for political expressions. This study employed a content analysis method to explore twenty-nine Thai political weblogs established during the period of Prime Minister Yingluck Shinawatra's administration. At the time, the most prominent Thai political weblogger was Nidhi Eawsriwong. Not surprisingly, the Pheu Thai Party and the Democrat Party were the most frequently appearing political parties in these weblog's posts. Most contents in these posts were related to government protesters by the People's Democratic Reform Committee (PDRC) and the coup d'état. The purposes of writing such weblogs were to express feelings and thoughts about Thai politics and to provide political information to the general public. The findings from this investigation revealed two significant uses of Thai political weblogs: the communication media for political expressions and viewpoints (a safe online space for political engagement and participation), and vital sources for Thai political information and news (social narratives).

Keywords: Thai political weblog, political communication, social media, political participation, political engagement

Received: May 19, 2020 **Accepted:** February 16, 2021

***Corresponding Author:** Sutthinan Chuenchom
https://orcid.org/0000-0003-1634-1961
E-mail: sutthinan@cmru.ac.th

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far 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.

© Sutthinan Chuenchom, 2021

1. INTRODUCTION

A blog (short for weblog), a user-generated, frequently updated website containing entries and comments concerning certain topics displayed in reverse chronological order, is an online medium for the purpose of digital communication. Weblogs have gained popularity due to the fact that they are easy to use, are virtually free of charge, and the author(s) of the blog (blogger/weblogger) can simultaneously publish content to various audiences. Weblogs are also a flexible and effective tool compared to other forms of publications.

As a channel for mass communication, weblogs are able to rapidly communicate and spread information to the public. This feature therefore provides an opportunity for various entities to freely express their thoughts and opinions, and to exchange information. Particularly in politics, weblogs have the potential to create a larger number of political spheres (online communities of people with similar interests). This enables the public to easily access and communicate political information through commonly available devices such as tablets and smartphones.

Political weblogging is a key medium for political communication (Yaakop et al., 2018). Therefore, the public has an opportunity to participate in democratic processes and information sharing. In both developed and developing countries, weblogs have been widely used as a channel for providing and obtaining political information or news, and for expressing political viewpoints. Yaakop et al. (2018) stated that political weblogs are widely used in order to share information and to allow citizens to freely and openly express their emotions and thoughts in cyberspace. Consequently, the use of digital platforms for political campaigning has grown globally since 2004 (Bankole et al., 2012). Several political parties use weblogs as an important part of their political campaigns.

Therefore, researchers have shown particular interest in such political weblogs in terms of their contents and effectiveness for civic and political purposes. Studies show that, in 2004 in the United States, weblogs were largely used for political campaigns (Williams et al., 2005). Bankole et al. (2012) studied the impact and use of weblogs in political elections in African countries in 2009. These studies revealed that even though the use of weblogs for campaigning in South Africa was at its early stages, a large number of webloggers, including African politicians, journalists, and media personnel, took advantage of the power of weblogs and were engaged in a high volume of political talk. Moreover, the use of weblogs in Asia was also being studied. Li (2008) found that several Chinese political weblogs were used to express dissatisfaction with the Chinese government. The influence of weblogs in politics indicates a growing trend in Hong Kong. Waikeung (2018) examined how political webloggers in Hong Kong used their weblogs to participate in the political scene during the 2014 'Umbrella Movement.' During this investigation, 960 political weblog posts were analyzed for content. The results of the study revealed that webloggers in Hong Kong used their platform as a form of a "soapbox," where they expressed their commentaries, opinions, and views in respect to recent public issues. As of this point however, there has been no academic research pertaining to the use of weblogs for political purposes in Thailand.

1.1. Emergence of Thai Weblogging

Thai weblogging started in 2009, and the number of weblogs in Thailand has been flourishing since. At that time, there were a large number of active weblogs in existence, for example, 2,703,115 blogs on www.bloggang.com, and 700,000 blogs on www.dek-d.com, respectively (Thammaneewong & Naktong, 2012). Another example is the Manager Online (www.mgronline.com), which is a news agency; however, they created their own weblog to allow ordinary people to become journalists. Similarly, the OK Nation Blog (www.oknation.net) promotes "citizen reporters" or "citizen journalists" by providing a hosting platform for everyone to report the news and to freely express their opinions online.

1.2. Need for the Study

Fundamentally, weblogging is very much related to democracy in order to facilitate political communication, discussions, and social mobilization, as well as collective civic actions (Sánchez-Villar, 2019), as this was also evident with webloggers in Hong Kong, who used weblogs to mobilize their readers to take political action (Waikeung, 2018). Obviously, weblogs provide the public with the opportunity and a platform to participate in democratic processes. As mentioned earlier, there has been no research in Thailand in regard to the use of weblogs for political purposes. Since Thailand is a democratic nation, the aim of this research is therefore to study the use of Thai political weblogs with four research questions: 1) What are the purposes of writing Thai political weblogs?; 2) Who are Thai political webloggers?; 3) Who and which political parties, political events, and organizations are related to Thai political weblog contents?; and finally, 4) How is the quality of content in Thai political weblogs? For the

purposes of this study, a content analysis was employed in order to examine Thai political weblogs in both Thai and English between 2011 and 2014, which coincides with the Prime Minister Yingluck Shinawatra administration. During this period, several remarkable political events had occurred, including the opposition to the Amnesty Act, the constitutional amendment, and the anti-government protests. Thai political weblog posts flourished significantly during this time, which provides a sufficient number of Thai political weblog posts in order to conduct a comprehensive analysis of the contents. Utilizing content analysis for the purpose of analyzing political weblogs provides valuable insight into changes and trends over time, and as well, it sheds light on patterns of interests and communication, preferences, attitudes, and behaviors, leading to conclusive outcomes pertaining to the sociology of the Thai political "blogosphere." The objective of this article is thus to analyze the Thai political sphere in order to better understand weblog practices in Thailand.

2. LITERATURE REVIEW

Social media have led to significant social changes by being a driving force behind social phenomena, especially when it comes to political issues (Laowishaya, 2012). Social media can rapidly reach the masses and consequently, political attitudes and participation are cultivated in apparent and visible fashion. Thus, social media have transformed political communication patterns, including political campaigns, ballots, and dialogues. Social media can therefore create larger political spheres, which encompass all dimensions and actors in politics such as politicians, political parties, movement groups, and independent organizations, as well as the public, who may all obtain political benefits from social media.

2.1. Weblogs

Basically, a weblog is a journal available on the Internet, which is maintained by a weblogger, who posts entries into the weblog; this activity is called "weblogging." Weblog entries are regularly updated in reverse chronological order and allow readers to leave comments. Webloggers can easily create weblogs and keep them current over time. The major characteristics of weblogs are that they are personalized, frequently updated, can link to other blogs and external websites, and can be easily published by a basic user interface. A weblog is a mix of text, images, and links to other blogs, web pages, or even other media objects such as graphics, multimedia, and Adobe PDFs. As a type of user-generated content media, webloggers may use their platform to write about any topic of interest. These interests can shift over time. Consequently, the topics of weblog posts range from individual diaries to political campaigns.

2.2. Political Weblogs

Over the past decades, weblogs have become computer-mediated tools for communication and for sharing and exchanging information and opinions (Herring et al., 2007). Politicians have employed weblogs to publicize their personal views, to discredit competitors, to campaign for their objectives, and to create political alliances, as well as to gain higher esteem in political arenas, and to convince audiences to follow and support their causes. Therefore as an end-user and a weblog follower, individuals can get more involved in the political process, and this provides the opportunity and an effective channel to receive political news directly from the source. Weblog followers can easily express their opinions about politics (as an online content creator), which may impose a significant impact on various dimensions of society. The nature of weblogs within the political arena has changed not only the way citizens produce, acquire, and process information, but it has also influenced the types of social movements and the manner in which individuals participate in political processes (Sánchez-Villar, 2019).

Therefore, many researchers have dedicated more time and energy to study political weblogs. McKenna and Pole (2008) found that webloggers generally write weblogs about policies of political parties. They often gather information from various sources on the Internet, and then use their skills to synthesize and analyze this information in their weblogs. Additionally, the public also use weblogs to communicate, exchange, and express their political viewpoints and opinions, as well as to address those issues relating to their local communities. Farrell and Drezner (2008) found that journalists often used an "A-list" political weblog as a news source. These A-list weblogs are those that provide the most relevant content or information, and as a result, they are the most popular weblogs to link to. Adamic and Glance (2005) analyzed the posts of 40 "Recommended weblogs" during the two months preceding the 2004 United States' presidential election. They found that there were not only links from these recommended weblogs to other weblogs, but also these weblogs were linked to various news articles. Right-leaning webloggers mentioned Democrats more often, while left-leaning webloggers mentioned Republicans more often.

Moreover, political webloggers, similar to mainstream journalists, supported their positions by criticizing those political figures whom they disliked.

In fact, weblogs are important for political media (Sánchez-Villar, 2019; Waikeung, 2018; Yaakop et al., 2018; Seper, 2004, cited in Lawson-Borders & Kirk, 2005). As mentioned earlier, the ever-increasing importance of weblogs in political scenes, such as in Hong Kong, is undeniable. Waikeung (2018) examined how webloggers in Hong Kong used their platform to influence politics. Waikeung (2018) divided the function of political weblogs into five categories: 1) informing readers about political issues or events; 2) acting as a "transmission belt": webloggers provide hyperlinks to other websites or quote other sources; 3) using their platform as a "soapbox": webloggers express their observations and opinions on political issues; 4) mobilizing political actions: webloggers appeal to their readers to take political actions; and finally, 5) acting as a conversation starter: webloggers pose questions and invite readers to provide feedback. Sánchez-Villar (2019) analyzed the role of political weblogs and posited weblogs as communication tools and citizen journalism. Similarly, Yaakop et al. (2018) analyzed the contents of 193 political weblogs during Malaysia's 2008 general election and concluded that political weblogs in Malaysia were indeed appropriate for delivering and disseminating information and for discussing issues that impacted the political landscape. Such scenarios have encouraged this study in order to understand the uses of weblogs in political arena in Thailand. The focus of this article is therefore on webloggers, the purposes for creating such weblogs, and the contents of such weblogs.

3. RESEARCH METHODOLOGY

This study focused on analyzing Thai political weblogs in accordance with Krippendorff (1989) content analysis process. The unit of analysis was Thai political weblog posts. The content analysis focused on three categories of weblogs based on purposes of writing: expressing feelings and thoughts, disseminating political information (to inform readers about political issues or events), and sharing knowledge and experience.

First, the researcher sought relevant weblogs both in Thailand and outside, and then created a list of Thai political weblogs. The weblogs' selection criteria were: 1) the weblog posts were published between 2011 and 2014; 2) contents of the posts were in both Thai and English; 3) the format of the contents could include text, images, or animation; and finally, 4) the posts were hosted on reliable and reputable weblogs. In total, there were 29 Thai political weblogs with 520 posts.

Subsequently, the researcher created a code book based on the components of the weblog and by assessing the quality of weblogs based on factors (Chuenchom, 2014) including: Authority (webloggers' name, expertise, qualifications, credentials, job positions, affiliations, contact information, relevance to topics), Credibility (number of posts, number of comments and originality), Content quality (word length, purposes of writing, related people and political parties, political events, organizations, reference/citation and subject headings).

Then, three librarians recorded the weblog posts' details and read the contents of each weblog's posts before categorizing them. Coding of weblog posts was processed independently based on the code book. Data was analyzed using descriptive statistics, including frequency and percentage, in order to answer the research questions. Finally, the librarians interpreted and drew inferences using descriptive statistics with the assistance of the researcher.

4. RESEARCH FINDINGS

The content analysis revealed that most Thai political weblogs were in Thai language and consisted of text and images.

4.1. Research Question 1: What Are the Purposes of Writing Thai Political Weblogs?

Thai political webloggers used their platform to criticize, discuss, or express their feelings and thoughts regarding Thai politics as shown in Table 1.

From Table 1, it is evident that the vast majority of Thai political weblog posts were used to express feelings and thoughts about Thai politics (86.35%), followed by those to disseminate information (24.04%). Only 3.84% of Thai political weblog posts were to share knowledge and experience.

Table 1. Purposes of Thai political weblogs

Purposes	f	%
Expressing feelings and thoughts	449	86.35
Disseminating political information	125	24.04
Sharing knowledge and experience	20	3.84

4.2. Research Question 2: Who Are Thai Political Webloggers?

The Thai political webloggers with the most number of posts were Nidhi Eawsriwong (10.58%), Asia Provocateur (9.23%), and Watcharaporn (6.73%), respectively. Professor Nidhi Eawsriwong is a well-known academic in political science at Chiang Mai University, Thailand. He is also a columnist, similar to Asia Provocateur and Watcharaporn, as shown in Table 2.

Most Thai political webloggers (75.19%) were anonymous; they did not provide their name, expertise, qualifications, credentials, job positions, affiliations, contact information, or webloggers' relevance to topic. Thus, most Thai political webloggers were at the lowest authority level.

In terms of credibility, four Thai political weblogs, those with the majority of posts, included Botkwamdee (50.00%), Asia Provocateur (10.77%), Chaoprayanews (10.38%), and Fhamon-konmong (8.46%), respectively, as shown in Table 3.

Most of the posts on these weblogs contained no comments (92.70%). If there were any comments, anonymity was often used, and the purposes of the comments were generally to support (63.64%) the weblogs' contents. Thai political posts had low originality since most Thai political weblog contents were reposted from other sources (58.08%). Matichon Online and Matichon Weekly were the Thai political weblogs with the largest number of posts as well as of those posts that were republished.

Table 2. Thai political webloggers

Thai political webloggers	f	%
Nidhi Eawsriwong	55	10.58
Asia Provocateur	48	9.23
Watcharaporn	35	6.73

Table 3. Thai political weblogs

Thai political weblogs	f	%
Botkwamdee (http://botkwamdee.blogspot.com)	260	50.00
Asia Provocateur (http://asiaprovocateur.blogspot.com)	56	10.77
Chaoprayanews (http://www.chaoprayanews.com)	54	10.38
Fhamon-konmong (http://fhamon-konmong.blogspot.com)	44	8.46

4.3. Research Question 3: Who and Which Political Parties, Political Events, and Organizations Are Related to Thai Political Weblog Contents?

From Table 4, the persons mentioned the most in the contents of Thai political weblog posts were Prime Minister Yingluck Shinawatra (24.49%), Abhisit Vejjajiva (the opposition leader) (13.15%), Thaksin Shinawatra (the former Prime Minister) (11.34%), Suthep Thaugsuban (the leader of the protest and the secretary-general of People's Democratic Reform Committee, PDRC) (9.75%), and General Prayut Chan-o-cha (Head of National Council for Peace and Order, NCPO) (5.44%), respectively. The political parties most related to the contents on Thai political weblog posts at that time were the Pheu Thai Party (Yingluck Shinawatra and Thaksin Shinawatra were in this) (67.35%) and Democrat Party (Suthep Thaugsuban and Abhisit Vejjajiva were in this) (28.91%), respectively.

4.4. Research Question 4: How Is the Quality of Contents in Thai Political Weblogs?

Most Thai political weblogs (21.15%) had 501-1,000 words, which means that Thai political weblogs (70.38%) often had low content quality since most Thai political weblogs were reposts with no reference and citation (98.08%).

From Table 5, the subject heading of most Thai political weblog posts pertained to Political violence (22.88%),

Table 4. Person and political parties to whom Thai political webloggers are related

Person	f	%
Prime Minister Yingluck Shinawatra	108	24.49
Abhisit Vejjajiva	58	13.15
Thaksin Shinawatra	50	11.34
Suthep Thaugsuban	43	9.75
General Prayut Chan-o-cha	24	5.44
Political parties	**f**	**%**
Pheu Thai Party	198	67.35
Democrat Party	85	28.91

Table 5. Thai political weblogs subject headings

Subject headings	f	%
Political violence	119	22.88
Government, Resistance to	107	20.58
Coups d'état	52	10.00
Coups d'état—Thailand	50	9.62

 https://doi.org/10.1633/JISTaP.2021.9.1.5

Government, Resistance to (20.58%), Coups d'état (10.00%), and Coups d'état—Thailand (9.62%). This corresponded to anti-government events that were largely discussed in Thai political weblogs at the time, which were initiated by the PDRC from August 2013 to May 2014 (3.46%), and the anti-coup (2.69%), respectively.

5. DISCUSSION

It appears that Thai political webloggers, in this research, frequently attempted to deliver political information. Professor Nidhi Eawsriwong empowered the Thai political sphere with the most weblog entries, which corresponds to Shao and Wang (2017)'s assumption that those with strong online presence enjoy higher social status in the real world. Moreover, using weblogs allows such actors and webloggers to gain benefits through the use of social networks (virtual communities and their relationships), and through an established reputation, as they can identify and record the most recent news and public interests. Most Thai political webloggers are scholars, which is different from the research findings in other countries. In those countries, a weblog is used as a digital platform for political campaigning, and most webloggers are politicians, representatives of the council, and journalists (Bankole et al., 2012; Li, 2008). It is important to note that those political webloggers who are actually scholars or experts in politics can help increase authority, content quality, and credibility of weblogs (Chuenchom, 2014).

Having said this, a large number of Thai political webloggers in this study were anonymous. Being anonymous is a unique feature in an online environment (Yun et al., 2016). Anonymous webloggers aim to either promote or discredit certain individuals or the government. However, the public perceives these weblogs to have higher level credibility. Weblog readers often believe the contents in such posts, as subsequent events are generally consistent with the weblog's projections or predictions. According to a survey by Technorati (Goldsborough, 2008), webloggers are often willing to tell stories about topics of interest, to share expertise and experiences with others, and to meet and interact with those who share these interests. Similarly, Thai political webloggers, in this study, showed great passion for information sharing. However, they identified themselves as anonymous writers or commentators, probably for their own safety, as it could be a risky proposition to criticize the government or political figures. As a general rule, Thai people refuse to express their opinions or speak publicly of others, which is rooted in the Thai culture. Indeed, under Thailand's censorship system, several Thai political weblogs have been shut down since they spoke of "untouched issues," or due to the fact that their thoughts were considered to be harmful to the security of the country. Likewise, webloggers in Malaysia appear to use nicknames and do not want to disclose their identities, perhaps due to the fact that they also want their identities to be hidden, for safety and other factors (Yaakop et al., 2018).

The content analysis of Thai political weblogs revealed that Thai political weblogs can be used as a source for Thai political news and information (Social Narrative), and furthermore, they may be used as an effective tool for communicating political expressions and viewpoints (a safe online space for Thai political engagement and participation).

5.1. Source of Thai Political News and Information (Social Narrative)

Thai political webloggers disseminate political information on their blogs, which can be seen as a convergence of important real-world political events and webloggers' interests. During the administration of Prime Minister Yingluck Shinawatra, the 36th Prime Minister of Thailand (5 August 2011-7 May 2014), several political phenomena occurred. Obviously, the contents on Thai political weblogs changed according to these political events. Hence, if one would sort through Thai political weblog posts of this period, it would become apparent that the timelines of such posts were consistent with those important Thai political events, 50 percent of which were re-posts from other sources. The most republished source was Matichon, a news agency with the main objective to convey news to the public. Obviously, Thai political weblogs serve as a tool for online journalism to communicate political news, situations, and phenomena in cyberspace. Huang and Lu (2017) explained that traditional mainstream media established new distribution channels by creating their own weblogs in order to expand their audiences, as weblogs became increasingly popular in people's life. Yun et al. (2016) emphasized that weblogs quite commonly embed news articles from traditional news organizations. Wallsten (2005) found that media coverage and weblog postings were also correlated. It is important to note that political weblogs facilitate and extend the reach of political news, which was previously provided by traditional mainstream media. Moreover, weblogs have become digital public spheres with the objective to improve citizens' ability to obtain and access daily political information and to

make political decisions. In fact, weblogs encourage civic engagement since they allow Thai people to freely access the media and to obtain information about Thai political events and issues. Additionally, Nelson et al. (2018) emphasized that weblogs could be used as an "alerter" of real-world political events that are generally not covered by the press or newspapers.

The Internet plays a major role in providing information. As a channel for continuous dissemination of information in unlimited quantities, and as a cost-effective non-stop option for transmitting information, the Internet offers individuals the right and opportunity to access information in an equitable manner. Accordingly, this helps to promote informed citizens, an important basis for strengthening democracy (Charoenwongsak, 2007). There is no doubt that journalism has been profoundly impacted by the existence of weblogs. The blogosphere can thus perform appropriate functions relating to participative journalism. The information functions and creation of public opinions make weblogs more pluralistic and qualitative (Sánchez-Villar, 2019). Indeed, weblogs have the capability to disseminate fruitful, relatively accessible, user-friendly information to a large number of readers. Also, they can provide feedback or comments (a two-way communication) rather than a space in an opinion column or a letter to the editor, as in conventional journalism. In addition, weblogs provide the opportunity to publish the latest information at a greater speed and offer the ability to link to various other information sources. Hence, weblogs can be "influencer platforms" for important citizen journalism, information nodes, and for content amplification. Emerging as news information sources and alternative tools for analysis of news stories, weblogs are "citizen-based" journals. Weblogs offer a wider range of news and information, and can help journalists to become aware of new stories, or to initiate stories.

Thai political weblogs are mostly written by scholars. Indeed, sharing political knowledge and experiences is one of the features of Thai political weblogs, which may offer a wider perspective and deeper insight on public interests. Typically, webloggers and weblog readers are dynamically involved in the process of information exchange. Not only do weblogs help to disseminate information, and access and manage content, but they also facilitate interaction, discussion, and collaboration. Moreover, weblogs can be useful for knowledge creation, as each weblog serves as a link that connects to different communities and helps in developing cross-disciplinary connections. Accordingly, analysis of the scope of knowledge in Thai political weblogs should be the focus of future research in order to discover whether or not Thai political weblogs are "political knowledge weblogs." Political knowledge weblogs are essential in people's daily lives in a knowledge-based society.

5.2. Communication Media for Thai Political Expressions and Viewpoints (A Safe Online Space for Thai Political Engagement and Participation)

The analysis of webloggers' writing purposes indicated that 86.35% of Thai political weblog posts pertained to political expression of thoughts and opinions relating to public concerns. Thai webloggers discussed Prime Minister Yingluck's government policies such as "the rice subsidy scheme" and the Constitutional amendments. The webloggers also raised issues of public concern including the anti-coup movement and the massive flooding in 2011. Obviously, weblogs are digital public spheres for expressing opinions; especially, weblogs provide a buffer zone for political critique. Yaakop et al. (2018) stated that people were able to pass on political messages without government control. Freedom of expression is a basic human right and therefore, Thai political weblogs are considered online political spaces and critical tools for everyone to promote political communication, engagement, and participation. Moreover, weblogs allow users to report political news with no boundaries as "citizen-based media" or citizen or grassroots reporters (Duong, 2017; Thammaneewong & Naktong, 2012), and to discuss and help to make decisions on current political issues. Online political engagement is likely to have the most immediate impact during a political crisis, and weblogging has emerged as one of the most important tools for mobilization and governance. In an information society, weblogs can be applied as means for information communication, and can function as a platform to support discussions in the political arena and among a variety of people with wide-ranging interests throughout socio-political systems (Nyblade et al., 2015; Sánchez-Villar, 2019; Sinpeng, 2014;). Webloggers act as information sources, so that readers can freely contribute contents to weblogs and discuss issues of specific interests, whereas the broader public recognizes weblogs as a space to voice their opinions and perspectives at no cost. Certainly, the blogosphere enhances the ability for political communication, which is in the dynamic nature of weblogs to perhaps positively impact democratization of the public sphere.

Noticeably, weblogs positively influence public politics. Political weblogs provide an accessible online platform for anyone to express their political opinions and feel-

ings, share experiences, and to start a conversation much more easily, no matter who they are and regardless of age, gender, and social status. The public has an opportunity to be both communicators and receivers, and to participate in democratic processes and information sharing. Application of technology can improve the quality of democratic systems, and foster civic and political engagement amongst citizens by reducing the gap between institutions and citizens, and by encouraging readers to take different types of political actions. Hence, political weblogs help to promote democracy. Remarkably, Thai political weblogs are an invaluable source of insight into people's opinions toward important Thai political events.

6. CONCLUSION

Thai political weblogs are accessible information channels that enable open participation in electronic public spheres. Weblogs facilitate political communication in the Thai political arena. Thai political weblogs are considered to be safe online spaces for expressing Thai political opinions and viewpoints, and a vital source for Thai political news and information. Moreover, political weblogs are digital public spheres, where the public can become reporters, communicators, or content and opinion creators. The research results support the idea that weblogs are one of the most effective communication tools in an information society. Indeed, political weblogs provide democratic discussion spaces, escalating a level of citizen engagement and participation regarding public issues. Any individual can be engaged in deeper and more open discussions of public interest and in exchanging information. Political weblogs undeniably provoke a political nature with a social reach and influence at no cost. Weblogs are public spheres where citizens, government, and information are interconnected. In order to become a more open and pluralistic information society, the credibility and quality of contents to inform and participate should be recognized, since online public opinions logically shape opinions in the real world.

Thai political weblogs are still in their early stages. Most Thai political webloggers are not politicians who use weblogs as a tool for public relations, for discrediting opponents, for self-campaigning and creating a positive attitude towards their own group, or for showcasing their movement to ensure people feel close to them. Thai political weblogs did not have any other forms of political communication, for example, political campaigns, ballots, and political dialogues during the period of Prime Minister Yingluck's government. It is apparent that the government, Members of Parliament, and other political representatives have not yet fully taken advantage of the power of weblogs to disseminate information and policies. However, political weblogs have already changed the pattern of public information. Moreover, news and information that people receive through these weblogs can be a foundation for social power and may change Thai politics in drastic ways. This study recommends that Thai political parties effectively employ weblogs as a tool for their public relations. Political entities can practice by building social relationships and communicating their messages using these tools in order to strengthen their bonds with the public and their constituents. Furthermore, Thai political parties and representatives can employ weblogs as a platform for the creation and dissemination of contents for a particular group of audiences. The use of technological applications can improve the quality of the Thai democratic system and foster greater Thai civic and political engagement in order to reduce the gap between institutions and citizens, resulting in a more empowered, healthy, and transparent democratic system.

In order to be good sources of information for public opinion, for public and private events, and for people in communities with similar interests, webloggers as well as readers should contribute value-added weblog contents by 1) providing links to other information sources; 2) building networks with real personal endorsements; and 3) increasing weblogs' credibility using critical and analytical skills to provide trustworthy, honest, and bias-free exchange. Thus, it could be concluded that weblogs are appropriate for political communication and transparent discussions of public issues and citizen journalism. They can create and provide more diverse information and open opinions, contributing to a wider perspective and deeper insight in political contexts.

The analysis of contents of Thai political weblogs enables information service providers to improve the quality of political information and to service users more effectively. In fact, lists of Thai political weblogs and Thai political webloggers are significant sources of political information. Also, the contents of Thai political weblog posts are like archives and record historically significant moments in Thai politics, which can be a notable Thai political learning source, especially for those captured during political crises in Thailand. Thai political weblog databases should be then developed. In addition, Thai political weblogs of past periods should be added to such databases. These databases may improve the quality of po-

litical information services. However, it is critical that the quality of Thai political weblogs be evaluated in terms of authority, credibility, or bias.

In order to better understand weblogs as a digital platform and information source for politics in a knowledge-based society, it is recommended to study the political information and knowledge in weblogs more intensively, and to pay close attention to biases and the actual exchanges on such political weblogs. Since weblogs positively contribute to information-citizen relationships, it is critical that research is done to determine the influence of weblogs on Thai politics, and furthermore, to assess sociopolitical impacts of weblogs on users' attitudes, participation, support for political events, and decision-making, which can be determined by examining various Thai political spheres.

CONFLICTS OF INTEREST

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

REFERENCES

Adamic, L. A., & Glance, N. (2005, August 21-25). The political blogosphere and the 2004 U.S. election: Divided they blog. In J. Adibi, M. Grobelnik, D. Mladenic, & P. Pantel (Eds.), *3rd International Workshop on Link Discovery: LinkKDD2005 in conjunction with 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 36-43). ACM. https://dl.acm.org/doi/10.1145/1134271.1134277.

Bankole, O. O., Chigona, W., & Bankole, F. O. (2012). The use of blogs for political campaigning in a developing country context: The case of South Africa. *TD: The Journal for Transdisciplinary Research in Southern Africa*, 8(1), 154-166. http://dspace.nwu.ac.za/handle/10394/6898

Charoenwongsak, K. (2007). Internet and democracy. Parliament Newsletter, 55(8), 117-131.

Chuenchom, S. (2014). User-centered evaluation of the quality of blogs. *Journal of Information Science*, 32(1), 58-71.

Duong, M. (2017). Blogging three ways in Vietnam's political blogosphere. *Contemporary Southeast Asia*, 39(2), 373-392. http://viet-studies.net/kinhte/Blogging3ways_CSE_August17.pdf

Farrell, H., & Drezner, D. W. (2008). The power and politics of blogs. *Public Choice*, 134(1-2), 15. https://doi.org/10.1007/s11127-007-9198-1.

Goldsborough, R. (2008). Report says Internet blogs growing in popularity, influence. *Community College Week*, 21(6), 16.

Herring, S. C., Scheidt, L. A., Kouper, I., & Wright, E. (2007). Longitudinal content analysis of blogs: 2003–2004. In M. Tremayne (Ed.), *Blogging, citizenship, and the future of media* (pp. 3-20). Routledge .

Huang, L., & Lu, W. (2017). Functions and roles of social media in media transformation in China: A case study of "@CCTV NEWS". *Telematics and Informatics*, 34(3), 774-785. https://doi.org/10.1016/j.tele.2016.05.015.

Krippendorff, K. (1989). Content analysis. In E. Barnouw, G. Gerbner, W. Schramm, T. L. Worth, & L. Gross (Eds.), *International encyclopedia of communication* (vol. 1, pp. 403-407). Oxford University Press. http://repository.upenn.edu/asc_papers/226.

Laowishaya, N. (2012). Social media in Thai political communication. *Executive Journal*, 32(1), 105-109.

Lawson-Borders, G., & Kirk, R. (2005). Blogs in campaign communication. *American Behavioral Scientist*, 49(4), 548-559. https://doi.org/10.1177%2F0002764205279425.

Li, L. (2008). *The social roles of Chinese political blogosphere in the age of "we media": A case study of Lianghui blogs*. (Master's thesis). Hong Kong Baptist University, Kowloon, Hong Kong.

McKenna, L., & Pole, A. (2008). What do bloggers do: An average day on an average political blog. *Public Choice*, 134(1-2), 97-108. https://doi.org/10.1007/s11127-007-9203-8.

Nelson, O., Loto, G., & Omojola, O. (2018). Blogging, civic engagement, and coverage of political conflict in Nigeria: A study of *nairaland.com*. *Kasetsart Journal of Social Sciences*, 39(2), 291-298. https://doi.org/10.1016/j.kjss.2018.05.007.

Nyblade, B., O'Mahony, A., & Sinpeng, A. (2015). Social media data and the dynamics of Thai protests. *Asian Journal of Social Science*, 43(5), 545-566. https://doi.org/10.1163/15685314-04305003.

Sánchez-Villar, J. M. (2019). The use of blogs as social media tools of political communication: Citizen journalism and public opinion 2.0. *Communication and Society*, 32(1), 39-55. https://doi.org/10.15581/003.32.1.39-55.

Sinpeng, A. (2014). Party-social movement coalition in Thailand's political conflict (2005–2011). In P. Liamputtong (Ed.), *Contemporary socio-cultural and political perspectives in Thailand* (pp. 157-168). Springer.

Shao, P., & Wang, Y. (2017). How does social media change Chinese political culture? The formation of fragmentized public sphere. *Telematics and Informatics*, 34(3), 694-704. https://doi.org/10.1016/j.tele.2016.05.018.

Thammaneewong, S., & Naktong, J. (2012). *Weblog as a creative learning media of new generation*. http://www.arts.su.ac.th/document/suwida/022.pdf.

Waikeung, T. (2018). Political participation by political bloggers in Hong Kong: A case study of the 2014 umbrella movement. *Journal of Current Chinese Affairs*, 47(1), 143-166.

Wallsten, K. (2005). *Political blogs and the bloggers who blog them: Is the political blogosphere and echo chamber?* Paper presented at the American Political Science Association Annual Meeting, Washington, D.C., USA.

Williams, A. P., Trammell, K. D., Postelnicu, M., Landreville, K. D., & Martin, J. D. (2005). Blogging and hyperlinking: Use of the web to enhance viability during the 2004 US campaign. *Journalism Studies*, 6(2), 177-186. https://doi.org/10.1080/14616700500057262.

Yaakop, M. R. M., Seman, A., Taib, R., Zawawi, H. Y., Jazimin, N. S., Dewi, S. A. A., & Saahar, S. (2018). Political blogging phenomenon to understand information behaviour and digital technology use. *International Journal of Academic Research in Business and Social Sciences*, 8(2), 149-157. https://doi.org/10.6007/IJARBSS/v8-i2/3863.

Yun, G. W., Park, S.-Y., & Lee, S. (2016). Inside the spiral: Hostile media, minority perception, and willingness to speak out on a weblog. *Computers in Human Behavior*, 62, 236-243. https://doi.org/10.1016/j.chb.2016.03.086.

APPENDIX. List of Thai Political Weblogs

1. Archive-it (http://wayback.archive-it.org)
2. Asia Provocateur (http://asiaprovocateur.blogspot.com)
3. Boringdays (http://www.boringdays.net)
4. Botkwamdee (http://botkwamdee.blogspot.com)
5. Chaoprayanews (http://www.chaoprayanews.com)
6. Downmerng (http://downmerng.blogspot.com)
7. Drphichai (http://drphichai.blogspot.com)
8. FACT – Freedom Against Censorship Thailand (https://facthai.wordpress.com)
9. Fhamon-konmong (http://fhamon-konmong.blogspot.com)
10. Hi-Thaksin (http://hi-thaksin.weebly.com)
11. ilaw (http://ilaw.or.th/node)
12. iTUKnowledge (http://ituknowledge.blogspot.com)
13. Kanj 2010 (http://kanj2010.blogspot.com)
14. Manaotoon (http://manaotoon.blogspot.com)
15. Matichon Online (www.matichon.co.th/)
16. Mblog (http://mblog.manager.co.th)
17. Mr-som (http://mr-som.blogspot.com)
18. Nation TV (http://oknation.nationtv.tv)
19. Newake (http://newake.blogspot.com)
20. OKnation blog (http://www.oknation.net)
21. Peera Panasupon (http://www.peerapanasupon.com)
22. Politics and Government (http://53011321092g12.blogspot.com)
23. Prasong.com (http://www.prasong.in.th)
24. Sermsak (https://sermsak.wordpress.com)
25. Somchai (https://somchaisr.wordpress.com)
26. Thai Society Today (http://luck507.blogspot.com)
27. Thamrongsak Petchlertanan (http://thamrongsakpetch.blogspot.com)
28. Warotah (http://warotah.blogspot.com)
29. Whatsoevermization (http://nandamization.blogspot.com)

 https://doi.org/10.1633/JISTaP.2021.9.1.5

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

RESEARCH PAPER
J Inf Sci Theory Pract 9(1): 79-89, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.1.6

Content Analysis of the Facebook Pages of Selected Academic Libraries in Vietnam

Duong Thi Phuong Chi*
Faculty of Philology, Perm State National Research University, Perm, Russia
Faculty of Library and Information Science, University of Social Sciences and Humanities (Vietnam National University - Ho Chi Minh City), Ho Chi Minh City, Vietnam
E-mail: duongthiphuongchi@gmail.com

ABSTRACT

This study explores Facebook use in Vietnamese academic libraries by analysing libraries' posts on their Facebook pages and library users' interaction with those posts. A total of 260 posts on four academic libraries' Facebook pages were examined using the content analysis method. The findings reveal that Facebook was mainly used to encourage reading and to transmit announcements. Most of the academic libraries published one post a week. The photo was the most frequent media type of libraries' posts and gained a higher level of interaction than other posts. According to the research results, the user engagement was low, and the user interaction with libraries' posts generally was in the form of reaction. The findings can help better understand Facebook use in Vietnamese academic libraries and may assist libraries in creating a plan for using Facebook more effectively.

Keywords: social networking sites, Facebook, academic libraries, library promotion, content analysis, Vietnam

Received: November 3, 2020 **Accepted:** February 1, 2021

***Corresponding Author:** Duong Thi Phuong Chi
https://orcid.org/0000-0002-2992-0957
E-mail: duongthiphuongchi@gmail.com

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far 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.

© Duong Thi Phuong Chi, 2021

1. INTRODUCTION

Social media is defined as "a group of Internet-based applications that built on the ideological and technological of Web 2.0" (Kaplan & Haenlein, 2010, p. 61), that allows its users to communicate, create, edit, and share contents (Lewis, 2010). The contents can be text, photo, video, or sound, or a mixture of all these (Scott, 2015). In this digital era, social media gives libraries a better chance to reach their users without distance and time limitations. King (2015) revealed that there are many benefits for libraries to use social media to understand the community, interact between libraries and users, receive more feedback from users, take advantage of mobile technology, and extend the reach of users. Kahn and Bhatti (2012) expressed that social media offers several opportunities for libraries to survive in a virtual environment.

Social media consists of many different platforms, such as social networking sites, blogs, discussion forums, video sharing, content sharing, social bookmarking, podcasts, and wikis (Sharma & Verma, 2018). Among them, social networking sites are much stronger than other platforms, which help people develop friendship environment and share information (Scott, 2015).

In recent years, Facebook has become a giant social networking site, with almost 2.8 billion monthly active users all over the world (Tankovska, 2021). There are over 55 million Facebook users in Vietnam as of April 2019, making it the seventh country in terms of Facebook audience size (Clement, 2019). Previous studies revealed that librarians have realized that social media, especially Facebook, has a positive effect. Therefore, just like other libraries abroad, Vietnamese academic libraries have started using Facebook as a powerful communication tool for libraries to transmit announcements, conduct online reference services, and introduce library resources to users. Although existing research mainly focused on the ways social media was used in Vietnamese libraries, the use of Facebook did not involve independent research. For all the reasons mentioned above, a study analysing Facebook use in Vietnamese libraries is necessary. In particular, this study aims to explore Facebook's current use in Vietnamese academic libraries. The primary research questions are as follows.

(1) What are the characteristics of academic libraries' posts on Facebook?
(2) How often do academic libraries post on their Facebook pages?
(3) How do users interact with libraries' posts?

This study's structure is as follows: first, a review of the literature is carried out. Then, the methodological approach is identified. Next, the analysis of collected data and discussion are presented, focusing on the main characteristics of libraries' posts, frequency of posts, and users' interactions with libraries' posts on Facebook. Finally, a conclusion is given.

2. LITERATURE REVIEW

2.1. Facebook

Facebook was first launched in February 2004 by students at Harvard University (Hall, 2012). Facebook is defined as a social networking site that allows its users to create a personal profile, maintain a friends list to whom messages can be sent, and join social networks organized by school, workplace, city, or region (Reitz, 2013). Access to Facebook is free of charge (Kohli et al., 2018). Facebook offers users various functions such as status updating, photo or video uploading, commenting on or reacting to posts, and sending private messages (Chan et al., 2020). Additionally, Facebook users can follow someone or a page and will then receive updates about them in their news feed (Wan, 2011). Facebook also provides an analytics section called Facebook Insights to measure usage and interaction (King, 2015). People have many reasons for using Facebook, including sharing and seeking information, communicating with family, friends, and colleagues, or leisure (Al-Busaidi, 2014; Marino et al., 2016). Facebook has become the largest social networking site today (Chugh & Ruhi, 2018; Shiau et al., 2018), with almost 1.85 billion daily active users during the fourth quarter of 2020 (Tankovska, 2021). The growing number of Facebook users shows that Facebook plays an increasingly important role in social life (Shiau et al., 2018). Divine et al. (2019) reported that 23% of Facebook users were young people aged 18-25 years, and 32% of undergraduate students spent more than four hours on Facebook daily.

2.2. Facebook Use in Academic Libraries

The research by Taylor and Francis Group (2016) reported that more than 70% of libraries use social media, particularly Facebook, which has become the most popular academic library platform (Cheng et al., 2020; Collins & Quan-Haase, 2012). Recent studies showed that academic libraries used Facebook for promoting library resources and services, reaching out to users, delivering online services, and enhancing teaching and learning (Cheng et al., 2020; Chugh & Ruhi, 2018; Vassilakaki &

 https://doi.org/10.1633/JISTaP.2021.9.1.6

Garoufallou, 2014; Wan, 2011). Specifically, several content analyses have been carried out on academic libraries' Facebook pages and have reflected various uses of Facebook by libraries in specific countries. For instance, studies examined academic libraries in the United States and found that libraries explored the applicability of Facebook as a tool to market services and make community connections using messages designed to pique the curiosity of users in an informal tone (Harrison et al., 2017; Phillips, 2011). Zorica et al. (2012) analysed the content of 91 Croatian libraries' Facebook pages and revealed that libraries mostly used Facebook for promotional purposes and posting links potentially interesting to users. However, a large number of library posts were considered "spam," which showed the lack of useful information. On the other hand, an analysis of Estonian academic libraries' Facebook pages found that Facebook was adapted mainly for announcements, with one post per week in a dull and formal form as in official documents. The library Facebook page could be considered as reduplicating a library website (Roos, 2013). At the same time, Tyson et al. (2019) said that Facebook was employed to deliver information literacy content to students.

Besides analysing the content, measuring the success of library posts has been discussed. Glazer (2012) stated that the level of engagement with posts is a significant indicator to evaluate Facebook use of libraries. Engagement is measured by the total number of likes, shares, and comments on posts (Al-Daihani & Abrahams, 2018; Gerolimos, 2011; King, 2015). Gerolimos (2011) investigated comments on the Facebook pages of American academic libraries and reported that 91% of posts did not have comments. Likewise, Al-Daihani and Abrahams (2018) analysed Facebook posts of academic libraries in the United States, Canada, the United Kingdom, and Australia to measure user engagement with library Facebook posts. Their findings showed a very low level of engagement, with less than a quarter of examined libraries gaining more than 2,000 post-related likes, and only seven libraries having more than 100 comments on posts.

2.3. Research Gap

Numerous studies on the use of Facebook have been conducted in recent years. However, they mainly mentioned libraries in the West. Thus, more case studies on the use of Facebook in various countries should be carried out to gain a comprehensive understanding of the Facebook phenomenon in libraries across the globe (Aharony, 2012; Chan et al., 2020). In Vietnam, few papers focused on the current trends and practical experiences of using social media in libraries have been found (Dương, 2019; Ngô, 2016; Trương, 2008). Although Facebook has been adapting to several Vietnamese academic libraries (Hồ, 2018), limited studies have been conducted on Facebook use. Therefore, the present research aims to fill a literature void by analysing academic libraries' posts and examining users' interaction with those posts.

3. METHODOLOGY

To know how academic libraries use Facebook, a content analysis of libraries' messages on their Facebook pages was conducted. In the past, researchers used a method called content analysis to identify the meaning of documents or other forms of communication (Allen & Reser, 1990). It has been applied to new technologies such as radio, television, and websites (Salinas, 2006). In the current study, content analysis is used to know how academic libraries are using Facebook, focusing on the categories of posts, the updates, and the degree of users' participation. The study developed the instrument to analyse libraries' Facebook pages based on the work of Nancy K. Phillips (Phillips, 2011). Posts on the chosen libraries' Facebook pages are in Vietnamese, but because the present study is carried out in English, it is necessary to translate posts into English.

This study is limited to the Facebook pages of four academic libraries in Ho Chi Minh City (Vietnam). The choice of academic libraries was made based on the fact that academic libraries are generally well-equipped with information resources and employees, compared to public libraries. Additionally, the criteria for choosing universities were: first, the universities are listed among the top universities located in Ho Chi Minh City (uniRank, 2020), and each university has a large population with more than 20,000 staff and students; next, selected libraries must have their own official Facebook pages. Academic libraries participating in this study include the Library of the University of Social Sciences and Humanities (USSH), Library of University of Science (US), Library of University of Technology and Education (UTE), and Library of Ton Duc Thang University (TDTU).

Only publicly available posts were used because of the unavailable data access from Facebook Insights. Considering the academic calendar of universities, the time of sampled posts ranged from September 1, 2019, to June 30, 2020 (one academic year), including messages posted by libraries or users. The data collection process comprised of four steps as following:

(1) Step one identified the library's Facebook pages. The appearance of the library Facebook page was embedded from the library website. Four library websites were scanned for the Facebook icon. The Facebook icon, which represents a symbol "f," was found on two library websites (US, TDTU). This icon immediately leads to the library Facebook page by clicking on it. Two other library Facebook pages were identified by asking librarians (USSH, UTE);
(2) Step two was to collect all posts appearing on each Facebook page between September 1, 2019, and June 30, 2020. The data were collected, including date of the post, content of the post, the number of reactions, the number of shares, and the number of comments for posts;
(3) In step three, each post on four libraries' Facebook pages were assigned categories;
(4) In the last step, the number of fans was recorded for each Facebook page as of June 30, 2020.

4. DATA ANALYSIS AND DISCUSSION

Although Facebook was launched in 2004 and became more popular in 2007, USSH was the first library which had its own Facebook page, in December 2012. It was noticed that all four libraries' Facebook pages provide general information about the library, such as opening hours, contact information, and library web page address.

4.1. Characteristic of Posts

4.1.1. Content of Posts

A total of 260 posts were collected from four libraries' Facebook pages. Separately, the appearance and percentage were calculated for each category, as shown in Table

Table 1. Categories of posts on four library Facebook pages from September 1, 2019, to June 30, 2020

Category	Definition	USSH	US	UTE	TDTU	Total (n=260)
Encourage reading	About author, book, reading	2	6	2	74	84 (32.31)
Announcement	Library hours, library policy, information about staff, the current status of service	18	8	9	39	74 (28.46)
Community building	Greeting, interacting with users by asking the question, inviting fans for joining competitions, promoting the image of the university, recognizing students, faculty	1	13	7	17	38 (14.62)
Promoting libraries	Highlight the attributes of libraries in general and the specific library	3	14	4	9	30 (11.54)
Event – library	Event organized or hosted by the library	4	1	9	9	23 (8.85)
Event – external	The event is not sponsored by or located at the library	5	4	0	8	17 (6.54)
Information resource	Identify database or information source available to users, or highlight a particular selection from an identified database or information resource, or an online resource that is on the world wide web	3	2	9	3	17 (6.54)
Instructional	Help users learn how to do something	0	1	5	8	14 (5.38)
General interest/ miscellaneous	Information that does not fit in other categories	1	4	0	5	10 (3.85)
Highlighting collection	Identify specific collection, excepting electronic databases or reference resources (categorized as Information Resource)	1	1	0	2	4 (1.54)
Other library services	Describe how the library can help users through specific services, including reference help, catalogs, inter-library loans, and others	1	0	1	2	4 (1.54)

Values are presented as number only or number (%).
USSH, Library of the University of Social Sciences and Humanities; US, Library of University of Science; UTE, Library of University of Technology and Education; TDTU, Library of Ton Duc Thang University.

 https://doi.org/10.1633/JISTaP.2021.9.1.6

1. It was feasible for a post to be allocated more than one category, as appropriate.

The results analysis revealed that academic libraries published various posts that showed who they are and what they do. Libraries also try to engage and connect with users via Facebook. Information that academic libraries delivered through their Facebook pages is as below.

4.1.1.1. Encourage Reading.

Table 1 shows that academic libraries emphasize reading as a principal value. A broad content category relates to encouraging reading with 84 posts, which collectively represent 32.31% of total posts. It was noticed that these messages were combined with other categories. For example, the author event promoted reading and provided a chance to introduce books in the library collection. Most encourage reading messages were posted by TDTU, representing over 88% (74/84) of total posts in this category. Moreover, TDTU also introduced exciting books in different ways, such as taking photos of highlighted important text in books, or making videos under two minutes to review the content of specific books, to encourage users to read them (Fig. 1).

4.1.1.2. Announcements.

Announcement posts gave users current information that they should know to use the library better. Announcements on four libraries' Facebook pages were 28.46% of the total. These announcements mainly focused on changes in library operations in the long term or temporary. Announcement posts also informed users when services were unavailable and when they were back in the process. Moreover, announcements were about changes in facilities, policies, and employees. Examples of announcement posts by sub-theme are given in Table 2.

From Table 2, some announcements were written relationally, for instance, "Many thanks for your patience!" to acknowledge that users may have felt inconvenience because the system was down.

Furthermore, announcement posts reflected the rhythms of the university calendar. The finding shows that the number of announcements in the beginning (September) and the end of the academic year (May and June) was higher than in other periods because most students used libraries at that time. These announcements related to registering new students, reserving a study room, and returning books. Besides this, Facebook was used to inform the immediacy of information, like changes in the weather,

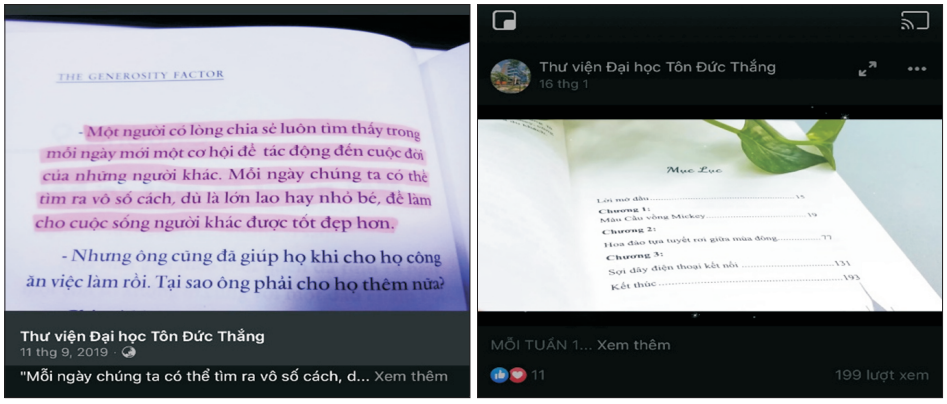


Fig. 1. Encourage reading posts on the Library of Ton Duc Thang University Facebook page.

Table 2. Examples of announcement posts

Sub-theme	Examples
Hours of operation	UTE posted on December 31, 2019: NEW YEAR'S HOLIDAY 2020! The Library will be closed on Wednesday (January 1, 2020)
Facility	US posted on September 5, 2019: Free wireless access is available within the Library.
Service availability	USSH posted on February 6, 2020: The network will be down from Fri, February 7, for maintenance. There will be no access to databases and the wireless network until Tue, February 11. Thank you for your patience!
Policies	TDTU posted on December 21, 2019: Please bring your card when visiting the Library!

UTE, Library of University of Technology and Education; US, Library of University of Science; USSH, Library of the University of Social Sciences and Humanities; TDTU, Library of Ton Duc Thang University.

which could help users. For example, "Do not forget your umbrella or raincoat!" was posted on TDTU's Facebook page status on May 25, 2020.

4.1.1.3. Community Building.

Another major category was how academic libraries shared with students, the university, and the local community to build strong relationships between them (Table 3).

Table 3 shows that academic libraries find ways to interact with their users casually via invitations, supportive comments, or direct questions. Several posts gave kindly wishes to users without mention of libraries. Like announcements, some greetings followed the university calendar, such as welcoming freshmen students at the beginning of the academic year and congratulating graduates at the end of the academic year. Besides this, libraries started direct interaction with followers by inviting them to participate in activities and asking questions related to libraries.

Libraries showed that their posts were not just about reading and books. They began to share messages about the university. There were posts on the achievements and contributions of the university, faculty, and staff. For example, US and TDTU shared news about the feelings of pride or creations of lecturers and researchers from their universities. Furthermore, academic libraries also shared content outside of the university. For instance, libraries promoted community programs to connect students to the community and push them to participate (i.e., US posted a message "Say "NO" to single-use plastic!" on September 19, 2019).

4.1.1.4. Library Promotion.

As mentioned earlier, libraries are taking advantage of opportunities to promote themselves, their resources, and services via Facebook with promotional posts, representing 11.54% of total posts. However, some of the posts which promoted libraries overlapped with community building and libraries' events categories.

4.1.1.5. Events, Information Resources, and Services.

Other content categories were related to posts about events, library collections, databases, and other services available to users. All four libraries' Facebook pages posted messages about libraries' events, which collectively represent 8.85% of total posts. Besides this, libraries were also informed about other events not located at or sponsored by libraries (6.54%). Also, libraries used Facebook to share instructional sessions (5.38%), to introduce electronic information resources available in libraries or on the Internet (6.54%), as well as to highlight impressive library collections (1.54%).

Via Facebook, academic libraries reminded users of reference services and reserves. Besides this, they collected feedback from users about library services. Table 4 gives examples of posts about events, information resources, and services by categories.

Facebook posts displayed various resources, services, and events that libraries offer to meet users' needs; therefore, this can help expand the brand of libraries in the digital era.

Generally, the content analysis shows what can be seen and inferred from posts on four libraries' Facebook pages. Results of the study demonstrated that Facebook offers a dynamic environment for libraries to communicate with users. There is general information about libraries on their Facebook pages. As expected, a significant portion of posts contains book recommendations, announcements, introduction-specific collections and databases, instructional sessions, and various events, which may influence users' perceptions about libraries' roles. The distribution of post categories varied across selected libraries. "Encourage reading" was the most frequent topic shared by TDTU, "Library promotion" was often posted on US's Facebook page, while "Announcements" posts were frequently posted by USSH and UTE. Several posts were written in an

Table 3. Examples of posts in the community building category

Sub-theme	Examples
Supportive, good wishes	USSH posted on September 8, 2019: Welcome Class of 2019!!!
Celebration	UTE posted on December 24, 2019: We wish you a Merry Christmas! (included photo of Christmas card)
Questions	TDTU posted on December 31, 2019: New Year is coming! Would you like the Library to do something? Please leave your comment below!
Society messages	US posted on September 19, 2019: Say "NO" to single-use plastic! (included photo)

USSH, Library of the University of Social Sciences and Humanities; UTE, Library of University of Technology and Education; TDTU, Library of Ton Duc Thang University; US, Library of University of Science.

 https://doi.org/10.1633/JISTaP.2021.9.1.6

Table 4. Examples of messages about events, information resources, and services

Category	Examples
Event – library	US posted on October 15, 2019: FREE FOOD AND DRINKS! Come to the Library to study or take a break.
Event – external	USSH posted on March 31, 2019: Play Online Quiz and Win Prizes. Springer Nature Publisher cooperates with VNU-HCM on organizing a "Springer Nature Online Quiz at VNU HCM." Quickly access the SpringerLink database to answer questions and bring back any exciting gifts, dear library users! (including the link to the event announcement)
Instructional	TDTU posted May 13, 2019: 17 Google search tips! You'll regret not knowing these tips sooner! (including instructional video)
Information resources	UTE posted on December 18, 2019: Find the best database for your research! Vietdata provides online, full-text access to publications on Business & Economics in Vietnam. Click on the link to access (including a link to the database)
Highlighting collection	TDTU posted on May 18, 2020: Celebrate Ho Chi Minh's 130th birthday. Discover research resources on Ho Chi Minh from the Library, including books and academic journals! (including photos of books and journals)

US, Library of University of Science; USSH, Library of the University of Social Sciences and Humanities; VNU-HCM, Vietnam National University – Ho Chi Minh City; TDTU, Library of Ton Duc Thang University; UTE, Library of University of Technology and Education.

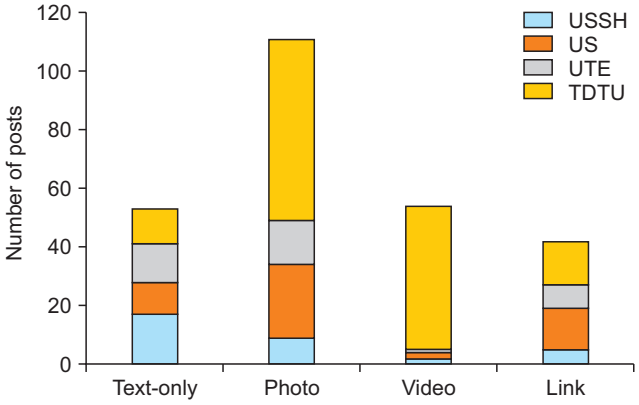


Fig. 2. Media types of posts. USSH, Library of the University of Social Sciences and Humanities; US, Library of University of Science; UTE, Library of University of Technology and Education; TDTU, Library of Ton Duc Thang University.

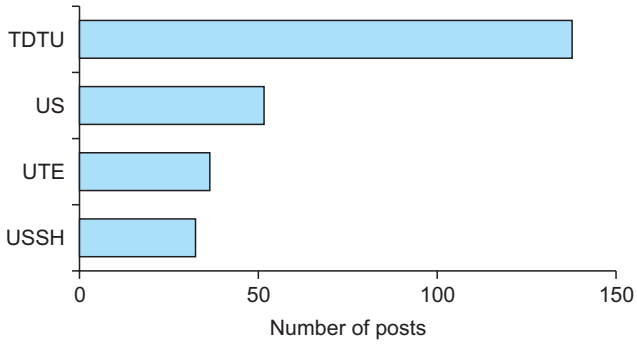


Fig. 3. Number of posts on library Facebook pages from September 1, 2019, to June 30, 2020. TDTU, Library of Ton Duc Thang University; US, Library of University of Science; UTE, Library of University of Technology and Education; USSH, Library of the University of Social Sciences and Humanities.

informal style by using punctuation such as capital letters and exclamation points.

4.1.2. Media Type of Posts

The media type was classified based on four types, including photo, video, link, and text-only. The media type of posts adopted by sampled libraries between September 1, 2019, and June 30, 2020, are shown in Fig. 2.

According to Fig. 2, the most frequent media type of posts was a photo (42.69%), followed by video (20.77%), text-only (20.38%), and link (16.16%) as less posted. The three libraries, including US, UTE, and TDTU, mostly used photo-based media for their posts, while USSH used text-only for posting. Notably, TDTU posted a large number of videos, with 35.51% of their total posts. Besides this, the US and TDTU libraries also preferred to share links. It seems that sharing content helps to maintain high posting frequency while saving time.

4.2. Frequency of Posts

The findings show that the frequency of posts on sampled libraries' Facebook pages was complicated. During the 43 weeks (September 1, 2019, to June 30, 2020), the number of Facebook posts ranged from 33 to 138, with USSH registering the least number of posts, 0.77 posts per week (n=33). Similarly, UTE also had a low number of posts, an average of 0.86 posts per week (n=37). On the other hand, TDTU had the most number of posts, which was equivalent to 3.2 posts per week (n=138), while the US updated its status by an average of 1.21 posts a week (n=52) (Fig. 3).

4.3. Interaction with Posts

Like other Web 2.0 tools, Facebook offers the option for interactivity between its users. People can make reactions (i.e., Like, Love, Haha, Wow, Sad, and Angry), "Share" or "Comment" on the post on the news feed. Facebook defines liking as a simple way to show enjoyment on posts or comments while reacting to specify a response (Facebook, 2020b). A "Comment" is the way people communicate on Facebook (Whitnah, 2010).

For the 260 posts coded in the current study, 100% of posts prompted reactions by fans. However, only 32.69% (n=85) of posts elicited comments, excluding comments of librarians. For these 85 messages, there were 363 comments (Table 5). It was noticed that the number of words was lower than the number of shares or reactions. Perhaps this was because a click on the response and share button is more comfortable than writing a comment.

Table 5 displays the number of fans for every library's Facebook page. Libraries' Facebook pages gained between 2,554 and 19,182 fans at the end of the data collection period. There was a difference in the number of fans between the chosen libraries. Among them, the UTE's Facebook page is the last one, with the least new fans.

The current study also calculated the engagement rate better to understand users' interaction on libraries' Facebook pages. According to Facebook, the engagement rate refers to the percentage of people who see a post that click, react to, comment on, and share it (Facebook, 2020a). Besides this, the engagement rate was influenced by the number of posts and fans, so that the average engagement rate was calculated as the sum of reactions, comments, and shares per post per fan (Lam et al., 2019).

$$Engagement\ rate = \frac{Number\ of\ (reactions+comments+shares)}{Number\ of\ Posts \times Total\ Fans\ on\ a\ Given\ Day} \times 100$$

According to Table 5, the engagement rate was different among the four libraries. US received the highest rate (1.5%), followed by UTE (1.06%), USSH (0.57%), and TDTU (0.1%). Surprisingly, TDTU had the highest number of posts and fans, but its engagement rate was the lowest. In contrast, the US did not update the post frequently; however, the US's Facebook page had the highest engagement rate.

Furthermore, the research finding examined which categories of posts attracted the most users.

According to Fig. 4, the "Promoting libraries" category gained the highest interactions from library users with 81 acts per post on average, following by the "Event – library" category with 60 acts, and the "Community building" category with over 43 acts. For instance, a promoting post of new facilities in the TDTU attracted 177 reactions, 30 comments, and 13 shares, or a post of a Christmas event at the TDTU also gained 219 reactions, 66 comments, and 52 shares (Fig. 5). The examples showed that library users were interested in new facilities that support their learning. Besides this, posts of library events in the form of hu-

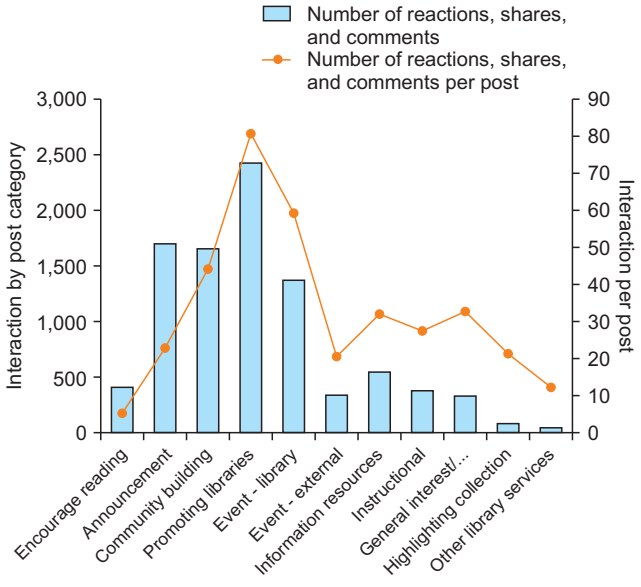


Fig. 4. Interactions of library users by post categories.

Table 5. Number of fans, reactions, comments, shares, and engagement rate from September 1, 2019, to June 30, 2020

Name of library	Number of fans as of June 30, 2020	Number of reaction	Number of shares	Number of comment	Average engagement rate (%)
USSH	10,340	1,674	200	84	0.57
US	3,979	2,842	120	149	1.5
UTE	2,554	861	110	34	1.06
TDTU	19,182	2,385	260	96	0.1

USSH, Library of the University of Social Sciences and Humanities; US, Library of University of Science; UTE, Library of University of Technology and Education; TDTU, Library of Ton Duc Thang University.

 https://doi.org/10.1633/JISTaP.2021.9.1.6



Fig. 5. Posts on the Library of Ton Duc Thang University Facebook page. (A) Post of "Promoting libraries" category. (B) Post of "Event – library" category.



Fig. 6. Posts on library Facebook pages. (A) Library of Ton Duc Thang University's post. (B) Library of the University of Social Sciences and Humanities's post.

morous and friendly approaches also impressed users.

Moreover, it is noticed that the interactions were different from the media type of post. While photo, video, and link posts attracted user attention mostly, text-only posts displayed a low level of attraction. For example, a post of TDTU with an amusing photo about the library's rule was more appealing and drew users' attention with 90 reactions, seven comments, and six shares. In contrast, although the posting of USSH was clear and easy to follow, a similar text-only message without a photo or video was not impressive. It has been observed that this post received only one "Like" from users (Fig. 6). Therefore, we believe that contents with a higher level of vividness are more attractive than plain text.

5. CONCLUSION

This study explores Facebook use in four academic libraries in Ho Chi Minh (Vietnam) by analysing libraries' posts and user interaction on libraries' Facebook pages. "Encourage reading" and "Announcements" were the most frequent content type posted by Vietnamese academic libraries on Facebook. The research results were quite the same as previous findings, such as finding that academic libraries' Facebook pages were not updated often, and content comprised of photos was the most favorite type of post. The current study is in line with the study of Al-Daihani and Abrahams (2018), which confirmed library users interacted most with photos. The number of comments was much lower than for other reactions, and the engagement rate was low among libraries studied. Generally, it can be concluded that selected libraries used Facebook as a tool for delivering information rather than as a venue for two-way interaction with library users. Therefore, academic libraries should upload content more frequently to make their Facebook pages more active. Posting with a higher level of vividness may be good practice for using Facebook. Furthermore, libraries should survey to know users' expectations and then diversify libraries' presences on Facebook accordingly.

The current study was limited to what can be seen on Facebook pages, so that there were unanswered questions without input from both librarians and library users. Besides this, the sample libraries are too tiny to generalize how Vietnamese academic libraries are using Facebook. For all these limitations, in-depth studies on factors influencing Facebook usage, managing Facebook's presence, user satisfaction from the library performance on Facebook, etc., are needed.

CONFLICTS OF INTEREST

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

REFERENCES

Aharony, N. (2012). Facebook use in libraries: An exploratory analysis. *Aslib Proceedings*, 64(4), 358-372. https://doi.org/10.1108/00012531211244725.

Al-Busaidi, K. A. (2014). SWOT of social networking sites for group work in government organizations: An exploratory Delphi study from IT managers' perspective. *VINE*, 44(1), 121-139. https://doi.org/10.1108/VINE-06-2012-0019.

Al-Daihani, S. M., & Abrahams, A. (2018). Analysis of academic libraries' Facebook posts: Text and data analytics. *The Journal of Academic Librarianship*, 44(2), 216-225. https://doi.org/10.1016/j.acalib.2018.02.004.

Allen, B., & Reser, D. (1990). Content analysis in library and information science research. *Library & Information Science Research*, 12(3), 251-262. https://eric.ed.gov/?id=EJ421718.

Chan, T. T. W., Lam, A. H. C., & Chiu, D. K. W. (2020). From Facebook to Instagram: Exploring user engagement in an academic library. *The Journal of Academic Librarianship*, 46(6), 102229. https://doi.org/10.1016/j.acalib.2020.102229.

Cheng, W. W. H., Lam, E. T. H., & Chiu, D. K. W. (2020). Social media as a platform in academic library marketing: A comparative study. *The Journal of Academic Librarianship*, 46(5), 102188. https://doi.org/10.1016/j.acalib.2020.102188.

Chugh, R., & Ruhi, U. (2018). Social media in higher education: A literature review of Facebook. *Education and Information Technologies*, 23(2), 605-616. https://doi.org/10.1007/s10639-017-9621-2.

Clement, J. (2019). *Leading countries based on number of Facebook users as of April 2019 (in millions)*. https://www.statista.com/statistics/268136/top-15-countries

Collins, G., & Quan-Haase, A. (2012). Social media and academic libraries: Current trends and future challenges. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1-4. https://doi.org/10.1002/meet.14504901272.

Divine, A., Watson, P. M., Baker, S., & Hall, C. R. (2019). Facebook, relatedness and exercise motivation in university students: A mixed methods investigation. *Computers in Human Behavior*, 91, 138-150. https://doi.org/10.1016/j.chb.2018.09.037.

Dương, T. P. C. (2019). Tiếp thị truyền thông xã hội trong hoạt động thư viện - thông tin [Social media marketing in library-information activities]. *Vietnam Library Journal*, 2, 19-24. https://nlv.gov.vn/tap-chi-thu-vien-viet-nam/tap-chi-thu-vien-viet-nam-so-2-2019-ra-ngay-15/03/2019.html.

Facebook (2020a). *About engagement rate ranking*. https://www.facebook.com/business/help/2351270371824148.

Facebook (2020b). *How do I react to a post or comment on Facebook?* https://www.facebook.com/help/933093216805622.

Gerolimos, M. (2011). Academic libraries on Facebook: An analysis of users' comments. *D-Lib Magazine*, 17(11/12), 5. http://doi.org/10.1045/november2011-gerolimos.

Glazer, H. (2012). "Likes" are lovely, but do they lead to more logins?: Developing metrics for academic libraries' Facebook pages. *College & Research Libraries News*, 73(1), 18-21. https://doi.org/10.5860/crln.73.1.8688.

Hall, M. (2012). *Facebook*. https://www.britannica.com/topic/Facebook.

Harrison, A., Burress, R., Velasquez, S., & Schreiner, L. (2017). Social media use in academic libraries: A phenomenological study. *The Journal of Academic Librarianship*, 43(3), 248-256. https://doi.org/10.1016/j.acalib.2017.02.014.

Hồ, T. N. (2018). *Ứng dụng truyền thông xã hội trong hoạt động thư viện Trường Đại học Khoa học Xã hội và Nhân văn Đại học Quốc gia Thành phố Hồ Chí Minh [Application of social media in library activities at the University of Social Sciences and Humanities]*. [Master's thesis]. Vietnam National University, Ho Chi Minh City, Ho Chi Minh City, Vietnam.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59-68. https://doi.org/10.1016/j.bushor.2009.09.003.

Khan, S. A., & Bhatti, R. (2012). Application of social media in marketing of library and information services: A case study from Pakistan. *Webology*, 9(1), Article 93. http://www.webology.org/2012/v9n1/a93.html.

King, D. L. (2015). *Managing your library's social media channels*. ALA TechSource.

Kohli, M. D., Daye, D., Towbin, A. J., Kotsenas, A. L., & Heilbrun, M. E. (2018). Social media tools for department and practice communication and branding in the digital age. *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, 38(6), 1773-1785.

Lam, E. T. H., Au, C. H., & Chiu, D. K. W. (2019). Analyzing the use of Facebook among university libraries in Hong Kong. *The Journal of Academic Librarianship*, 45(3), 175-183. https://doi.org/10.1016/j.acalib.2019.02.007.

Lewis, B. K. (2010). Social media and strategic communication: Attitudes and perceptions among college students. *Public Relations Journal*, 4(3), 1-23. https://prjournal.instituteforpr.org/wp-content/uploads/2010Lewis.pdf.

Marino, C., Vieno, A., Pastore, M., Albery, I. P., Frings, D., & Spada, M. M. (2016). Modeling the contribution of personality, social identity and social norms to problematic Face-

book use in adolescents. *Addictive behaviors*, 63, 51-56.

Ngô, T. T. (2016). Ứng dụng marketing truyền thông xã hội trong cơ quan thông tin - thư viện [Application of social media marketing at information – library organizations]. *Tạp chí Thông tin và Tư liệu*, 5, 13-19. https://vjol.info.vn/index.php/VJIAD/article/view/25821.

Phillips, N. K. (2011). Academic library use of Facebook: Building relationships with students. *The Journal of Academic Librarianship*, 37(6), 512-522. https://doi.org/10.1016/j.acalib.2011.07.008.

Reitz, J. M. (2013). *Online dictionary for library and information science: ODLIS*. https://products.abc-clio.com/ODLIS/odlis_f.aspx.

Roos, J. (2013, January 23-25). Why university libraries don't trust Facebook marketing? In Z. Taşkın, T. Çakmak, & D. Doğan (Eds.), *The 21st International BOBCATSSS Conference* (pp. 160-165). Hacettepe University Department of Information Management.

Salinas, R. (2016). A content analysis of Latina Web content. *Library & Information Science Research*, 28(2), 297-324. https://doi.org/10.1016/j.lisr.2006.03.007.

Scott, D. M. (2015). *The new rules of marketing & PR: How to use social media, online video, mobile applications, blogs, news releases, and viral marketing to reach buyers directly*. Wiley.

Sharma, S., & Verma, H. V. (2018). Social media marketing: Evolution and change. In G. Heggde, & G. Shainesh (Eds.), *Social Media Marketing* (pp. 19-36). Palgrave Macmillan.

Shiau, W. L., Dwivedi, Y. K., & Lai, H. H. (2018). Examining the core knowledge on Facebook. *International Journal of Information Management*, 43, 52-63. https://doi.org/10.1016/j.ijinfomgt.2018.06.006.

Tankovska, H. (2021). *Number of daily active Facebook users worldwide as of 4th quarter 2020 (in millions)*. https://www.statista.com/statistics/346167/facebook-global-dau/.

Taylor & Francis Group (2016). *Use of social media by the library*. https://librarianresources.taylorandfrancis.com/white_paper/use-of-social-media-by-the-library/.

Trương, Đ. L. (2008). Sử dụng blog để phổ biến dịch vụ thông tin thư viện [Using blogs to disseminate library and information services]. *Vietnam Library Journal*, 4, 16-24. https://nlv.gov.vn/nghiep-vu-thu-vien/su-dung-blog-de-pho-bien-dich-vu-thong-tin-thu-vien.html.

Tyson, A. F., Angelo, A., McElwaine, B., & Tauro, K. (2019). Delivering information literacy via Facebook: Here comes the spinach! *Evidence Based Library and Information Practice*, 14(2), 33-50. https://doi.org/10.18438/eblip29532.

uniRank. (2020). *Top Universities in Vietnam: 2020 Vietnamese university ranking*. https://www.4icu.org/vn/.

Vassilakaki, E., & Garoufallou, E. (2014). The impact of Facebook on libraries and librarians: A review of the literature. *Program: Electronic Library and Information Systems*, 48(3), 226-245. https://doi.org/10.1108/PROG-03-2013-0011.

Wan, G. (2011). How academic libraries reach users on Facebook. *College & Undergraduate Libraries*, 18(4), 307-318. https://doi.org/10.1080/10691316.2011.624944.

Whitnah, T. (2010). *I like your comment*. https://www.facebook.com/notes/facebook-app/i-like-your-comment/399440987130.

Zorica, M. B., Ivanjko, T., & Benčec, M. (2012, May 21-25). Croatian libraries on Facebook - content analysis. In P. Biljanovic (Ed.), *35th International Convention MIPRO 2012* (pp. 1130-1134). Institute of Electrical and Electronics Engineers.

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? Ji-Young Kim (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 Article Contribution Management System (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

With the KISTI's Open Access Policy, authors can choose open access and retain their copyright or opt for the normal publication process with a copyright transfer. If authors choose open access, their manuscripts become freely available to public under Creative Commons license. Open access articles are automatically archived in the KISTI's open access repository (KPubS, www.kpubs.org). If authors do not choose open access, access to their articles will be restricted to journal users.

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

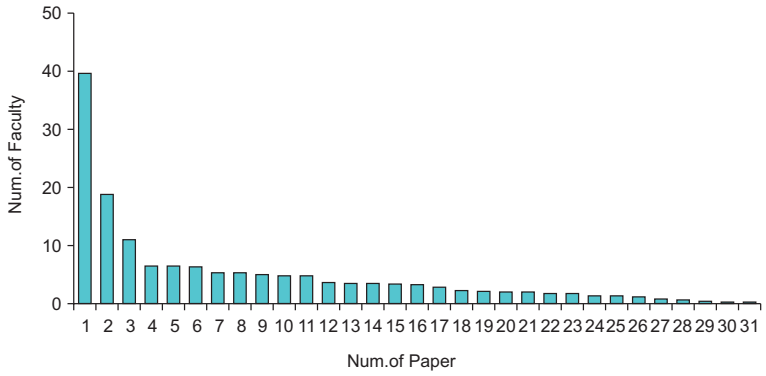


Fig. 1. Distribution of authors over publication count.

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

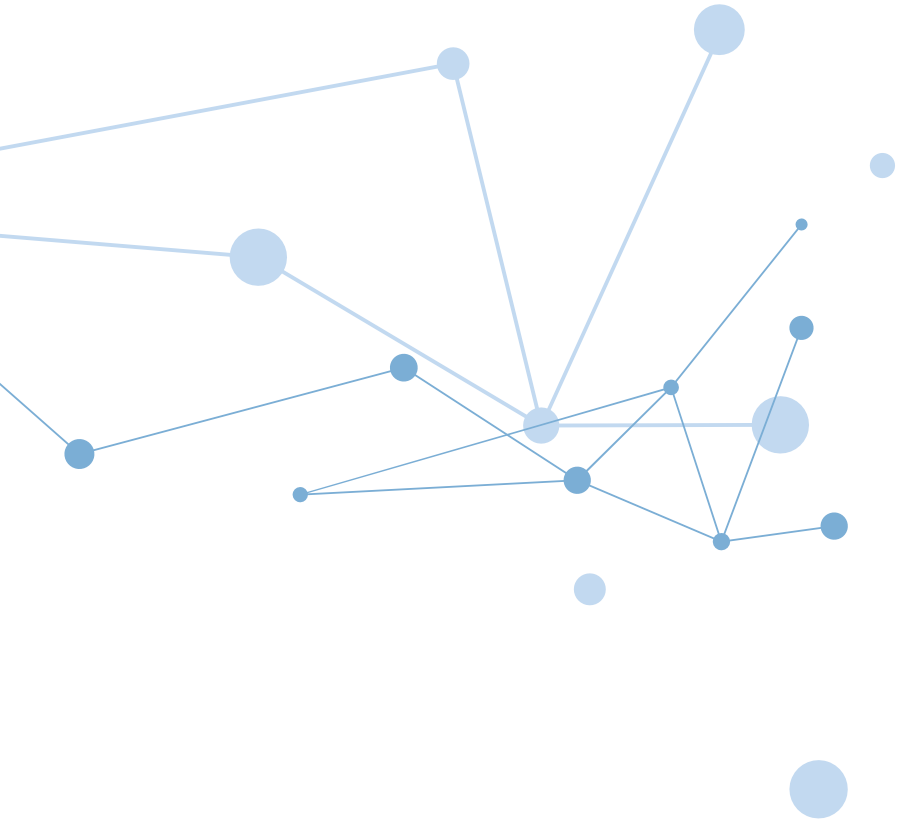
- Author, A., & Author, B. (Year). Chapter title. In A. Editor, B. Editor, & C. Editor (Eds.), *Book title* (pp. xx-xx). Publisher Name.
- Smith, K. & Brown, M. (2012). Author impact factor by weighted citation counts. In G. Martin (Ed.), *Bibliometric approach to quality assessment* (pp. 101-121). Springer.

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