Knowledge discovery of scholarly publications on misinformation on social media: A text mining approach

Prasadi Kanchana Jayasekara[1]

Abstract

Social media is a remarkable outcome of Web 2.0 technology, which is very popular among the Internet users. The general public is using social media as a communication media to fulfil their information requirements on various occasions such as disaster communication, health communication, marketing products and services and political campaigns. However, the openness of social media provides a great platform for misinformation sharing which is a very common problem in social media communication. Hence, the main objectives of this study are to analyse the publication year and total citation count of publications on misinformation on social media and to identify the main disciplines of misinformation studies on social media using the text mining technique. The primary data of this study were extracted from the Web of Science database using the keywords; social media, misinformation, disinformation and fake news on 16th April 2020. The WoS provided 62 search results and all 62 articles were considered in this study. The title of the article, journal, published year, total citation, abstract, author keywords, keywords plus, Web of Science categories, and research areas were extracted from WoS database. The text mining was done manually. According to the results, scholarly publications on misinformation on social media were first published in the year 2012. Scholarly publications were categorized into 10 main categories; Information, Media, Medical information, Social Science, Communication, Health information, Computer science, Other Sciences, Engineering and Management and Finance. The medical information subject area is covering vast varieties of research areas than the other main subject areas.

Keywords: Disinformation, Fake news, Text mining

[1]*Senior Assistant Librarian, Main Library, University of Ruhuna, Sri Lanka*
email: kanchana@lib.ruh.ac.lk

Introduction

Social media is the new phenomenal rise among Internet users. It has gained significant popularity in recent years. Social media are "websites and application that enable users to create and share content or to participate in social networking" (Oxford Dictionaries, 2018). That provides plenty of opportunities for the public to generate and share information with each other anywhere, anytime on a web-based interface. The information created by the public is called as User Generated Content (UGC).

Before the evolution of ICT people used an oral communication method called Word-of-Mouth (WOM) to fulfil their information need. However, the progression of Web 2.0 technologies and social media has shifted WOM to electronic Word-of –Mouth (eWOM). eWOM can be described as "any positive or negative statement made by potential, actual or former customer about a product or company, which is made available to a multitude of people and institutions via the Internet" (Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004). There is a potential of rapid spreading of negative eWOM than positive eWOM (Kok et al., 2014).

User generated content is the most powerful and popular form of eWOM. It can be an online comment, photograph, video, and any form of content created and uploaded by the general public in a digital platform (Webopedia, 2018). The social media act as a platform which provides opportunities to create and share UGC (A. M. Kaplan & Haenlein, 2010). Among all social media platforms, social networking sites are considered as one of the key platforms used to create user generated content (Bunce, Partridge, & Davis, 2012; Xiang & Gretzel, 2010).

Social media has changed the way of communication from traditional to modern. The general public is using social media in disaster communication (Bunce et al., 2012; Dougall, Horsley, & McLisky, 2008); health communication (Bannor, Asare, & Bawole, 2017; Henderson et al., 2017; Lim, 2016); marketing products and services (Moro & Rita, 2018; Pantano & Di Pietro, 2013); and for political campaigns (Udanor, Aneke, & Ogbuokiri, 2016; Vraga, 2016). However, because of the openness of social media platforms

information overloading and misinformation sharing are very common problems. Generally, individuals who seek information via online media, do not spend time to evaluate online information (Metzger, 2007). Misinformation is "false or inaccurate information which is unintentionally spread" (Wu, Morstatter, Hu, & Liu, 2016, p. 124). It may be disinformation, rumour, urban legend, spam, or troll (Wu, Morstatter, & Liu, 2016). Disinformation is "intentionally false and deliberately spread" (Wu, Morstatter, Hu, et al., 2016, p. 124).

Text mining is a popular technique used to identify research trends in different fields (Jayasekara & Abu, 2018; Tirunillai & Tellis, 2014; Zawacki-Richter & Naidu, 2016). Text mining is a good tool for obtaining a big picture about a particular topic. This study was carried out on identifying main and sub disciplines of scholarly publications focusing on misinformation on social media using text mining technique, which will provide an overview of scholarly publications in this discipline. Moreover, according to the literature survey, no prior studies have conducted focusing on scholarly publications on misinformation on social media using text mining technique.

Objectives

The main objectives of this study are;

1) To analyse the publication year and total citation count of publications on misinformation on social media.
2) To identify the main and sub disciplines of scholarly publications on misinformation on social media using text mining technique.

Literature review

Misinformation will cause unnecessary fear and mislead people. The concept of information credibility contains the credibility of the message, source and media (Metzger, Flanagin, Eyal, Lemus, & Mccann, 2003). However, due to the openness, measuring the credibility of social media is not an easy task. Moreover, the Internet provides vast opportunities and platforms that individuals can create and share information locally and globally. Therefore,

many studies covering various subject disciplines have identified studies on misinformation on social media as a fertile ground.

Text mining is an information extraction method which is also known as data mining or text analysis. Data mining is "the process of discovering interesting patterns and knowledge from large amounts of data" (Han, Kamber, & Pei, 2012, p. 45). There are so many researchers utilized text mining techniques to identify research trends in different fields, such as marketing (Tirunillai & Tellis, 2014), education (Zawacki-Richter & Naidu, 2016), data mining (Jayasekara & Abu, 2018) and statistics (Kaplan, Haudek, Ha, Rogness, & Fisher, 2014).

Web of Science is a citation database which covers ten indexed that managed by Clarivate Analytics. The WoS database search results provide mainly the following information; the title of the article, journal, published year, total citation, abstract, author keywords, keywords plus, Web of Science categories, and research areas. Here, author keywords are keywords provided by the authors of the article. However, keywords plus are automatically generated index terms based on the titles of the cited articles (Clarivate Analytics, 2020a). Generally, WoS is assigning subject category/categories to the scholarly publications included in their database which is called as Web of Science categories (Clarivate Analytics, 2020c). Currently, WoS has around 250 WoS Categories. Research areas are another indexing system used by WoS. WoS database categories scholarly publications into five main categories; Arts& Humanities, Life Sciences and Biomedicine, Physical Sciences, Social Sciences and Technology (Clarivate Analytics, 2020b). The main aim of assigning all of these keywords types is to facilitate efficient and effective information searching, retrieving and analysing facilities to their users.

Methodology

Data collection

The primary data of this study were extracted using the Web of Science (WoS) database. The database was searched using following keywords and Boolean operators to broaden the search area, (((Social media AND misinformation) OR (Social Media AND Disinformation)) OR (Social media AND fake news)).

Data were collected from the WoS core collection indexes including Science Citation Index Expanded (SCI-Expanded), Social Science Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI). The search strategy in Figure 1 was used to search the WoS database.

TITLE: (((Social media AND misinformation) OR (Social Media AND Disinformation)) OR (Social media AND fake news))
Timespan=All years. Indexes=SCI-EXPANDED, SSCI, A&HCI.

Figure 1 Search Strategy used to search the Web of Science database

The WoS database was accessed on 16th April 2020. The WoS provided 62 search results published within the year 1989 to 2020 for the search strategy used in this study. Hence, all 62 articles were considered in this study. The WoS database was used to extract the title of the article, journal, published year, total citation, abstract, author keywords, keywords plus, Web of Science categories, and research areas.

Data Analysis

The text mining was done manually. The researcher thoroughly examined the title, abstract, author keywords, keywords plus, Web of Science categories, and research areas of individual articles. Based on the content of the articles and with the help of the tags (author keywords, keywords plus, web of science categories, and research areas) provided by WoS database, the relevant main categories and sub categories were created manually by the researcher. Then each of the sixty two articles were separately examined for their content and assigned main categories and sub categories for each article.

Limitations

There are some limitations to this study. The usage of WoS database to extract data is an easy method. However, it only provides scholarly publications indexed in the WoS. There may be more publications published focusing misinformation on social media which are not indexed in WoS. Hence, the used data set may be incomplete.

Results

Figure 2 demonstrates the year wise distribution of scholarly publications on misinformation on social media. As revealed by the Figure 2, scholarly publications on misinformation on social media have a short history. The first publication was published in the year 2012. The highest number of publications were published in the year 2019 (26 publications) followed by the year 2018 with 12 publications. In the year 2020, 11 publications were published. However, the data was collected on 16th April 2020. Hence, the number of publications within the year 2020 may increase in the future.

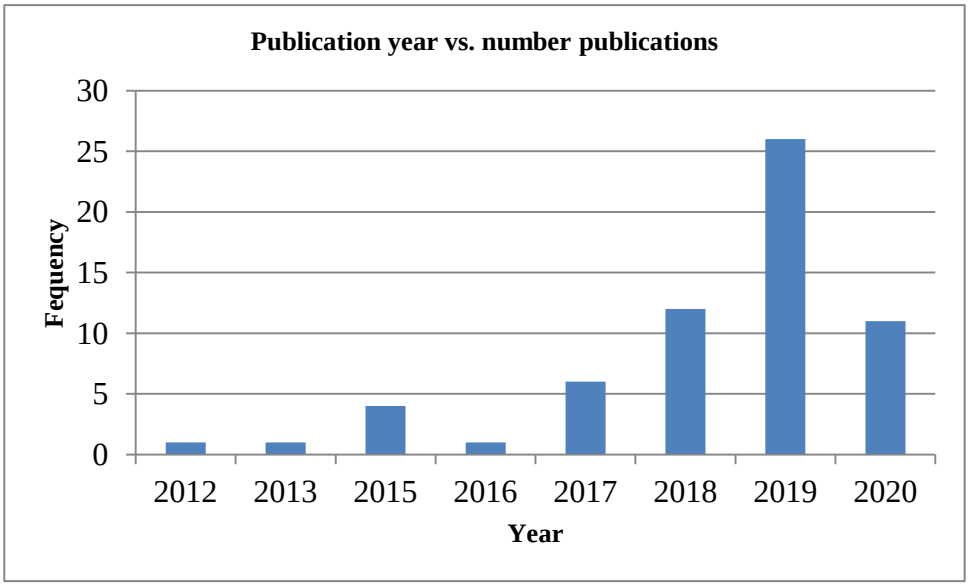

Figure 2 Publication year vs. the number of publications

The Figure 3 illustrates the total citation count received by every scholarly publication. The total citation counts of considered publications were vary from 74 to zero. The article titled "In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media" published in Journal of Communication, received the highest total citation count (74) followed by the article titled "Social Media and Fake News in the 2016 Election" published in Journal of Economic Perspectives with 39 total citations. A publication titled "Addressing Health-Related Misinformation on Social Media" published in JAMA-Journal of The American Medical

Association received third place with 36 total citation count. There were 33 publications with zero citations.

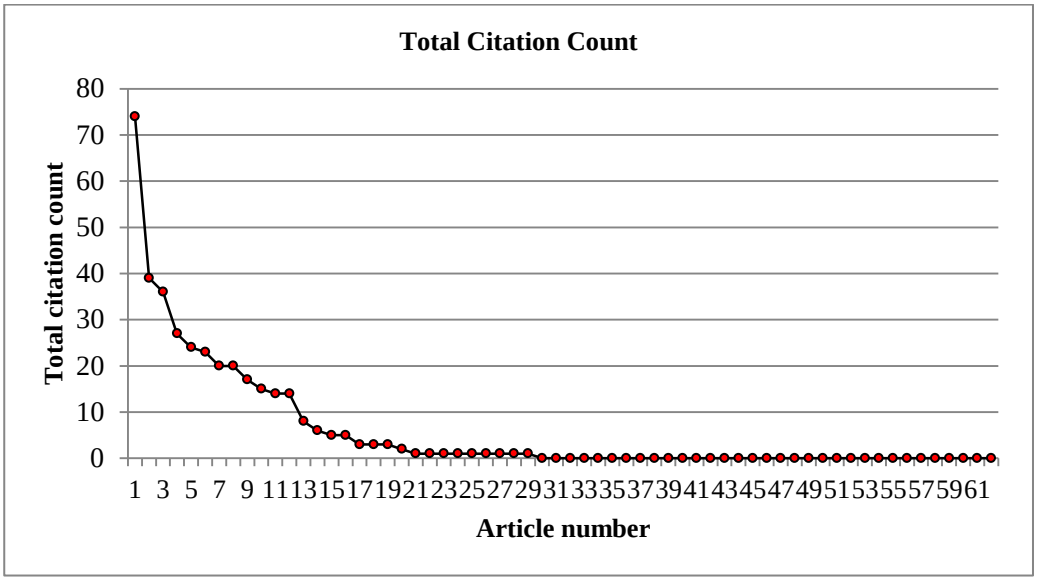

Figure 3 Total Citation Count

Table 1 show the main categories identified in this study. As illustrated in the Table 1, scholarly publications were categorized into 10 main categories based on the content such as Information, Media, Medical information, Social Science, Communication, Health information, Computer science, Other Sciences, Engineering and Management and Finance. According to the Table 1, 'Information' was the most popular subject category (54 publications) followed by 'Media' with 44 publications and 'Medical information' with 35 publications. However, there were few studies focused on 'Engineering' and 'Management and Finance' subject categories.

Table 1 Composition of main categories

Main Category	No of publications
Information	54
Media	44
Medical information	35
Social Science	29

Communication	16
Health Information	15
Computer science	9
Other Sciences	7
Engineering	5
Management and Finance	5

Table 2 illustrates the composition of subcategories. Subcategories show the research areas covered by scholarly publications under the main category. The values inside the parenthesis demonstrate the number of articles that covered the particular research area.

Table 2 Composition of sub categories

Main Categories	**Sub Categories**
Information	Misinformation (35); Sharing Behaviour (5); Fact-checking (5); Disinformation (5); Rumours (3); Information Literacy (1)
Media	Fake News (31); News media (5); Journalism (3); Telecommunications (2); Media trust (2); Film, Radio, Television (1)
Medical information	General & Internal Medicine (7); Medicine, General & Internal (6); Vaccination (5); Ebola (3); Zika Virus (2); Critical Care Medicine (1); Infectious Diseases (1); Medical Informatics (1); Peripheral Vascular Disease (1); Pediatrics (1); Tropical Medicine (2); Urology & Nephrology (1); Cardiovascular System & Cardiology (1); Dentistry, Oral Surgery & Medicine (1); Emergency Medicine (1); Medical Ethics (1)
Social science	Information Science & Library Science (8); Government & Law (5): Psychology, Multidisciplinary (4); Political Science (4); Social Sciences - Other Topics (2); Social Sciences, Biomedical (2); Psychology, Experimental (2); Social Sciences, Interdisciplinary (1); Medical Ethics, Social Issues (1)

Communication	Communication (15); Sociology (1)
Health information	Public, Environmental & Occupational Health (7); Health Care Sciences & Services (4); Health Policy & Services (3); Health Literacy (1)
Computer science	Information Systems (7); Interdisciplinary Applications (1); Software Engineering (1)
Other sciences	Physics, Multidisciplinary (2); Multidisciplinary Sciences (2); Science & Technology - Other Topics (2); Veterinary Sciences (1)
Engineering	Electrical & Electronic (4); Biomedical (1)
Management and Finance	Management (3); Economics (1); Business (1)

According to the Table 2, scholarly publications focused on 'Information' main category consisted of six sub categories. Those studies were focused on misinformation, information sharing behaviour, fact-checking, and disinformation, rumours and information literacy. Furthermore, Table 2 revealed that the 'Medical information' main category has the highest number of sub categories followed by 'Social Science' main category. Accordingly, it shows that publications related to the medical information subject area are covering vast varieties of research areas than the other main subject areas.

Discussion and Conclusion

Social media platforms are great invention of Web 2.0 technologies because it allows human to maintain their network with each other. However, due to the openness and freedom for the creation and sharing of information, it is using as a platform to share misinformation throughout the world. According to the results of this study, misinformation on social media is a new research area with a short history. Scholarly publications on misinformation on social media were covering 10 main subject areas; Information, Media, Medical information, Social Science, Communication, Health information, Computer science, Other Sciences, Engineering and Management and Finance. 'Information' is the most popular subject category followed by 'Media' and 'Medical information'. 'Medical information' main category has the highest number of sub categories followed by 'Social Science' main category.

This study is providing an overview of research areas covered on misinformation on social media. As a result, researchers can use this study to identify the research areas that they can focus on when conducting studies on misinformation on social media in the future. Moreover, this article provided an up-to-date knowledge of misinformation on social media research in the world. Researchers can use the same methodology used in this study to discover research trends in other areas.

References

Bannor, R., Asare, A. K., & Bawole, J. N. (2017). Effectiveness of social media for communicating health messages in Ghana. *Health Education, 117*(4), 342–371. https://doi.org/10.1108/HE-06-2016-0024

Bunce, S., Partridge, H., & Davis, K. (2012). Exploring information experience using social media during the 2011 Queensland Floods: a pilot study. *The Australian Library Journal*, *61*(1), 34–45. https://doi.org/10.1080/00049670.2012.10722300

Clarivate Analytics. (2020a). Keywords. Retrieved February 5, 2020, from Web of Science Core Collection Help website: https://images.webofknowledge.com/images/help/WOS/hp_full_record.html#dsy1028-TRS_keywords_plus

Clarivate Analytics. (2020b). Research areas. Retrieved February 5, 2020, from Web of Science Core Collection Help website: http://images.webofknowledge.com/WOKRS5251R3/help/WOS/hp_research_areas_easca.html

Clarivate Analytics. (2020c). Web of Science Categories. Retrieved February 5, 2020, from Web of Science Core Collection Help website: https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html

Dougall, E. K., Horsley, J. S., & McLisky, C. (2008). Disaster Communication: Lessons from Indonesia. *International Journal of Strategic Communication*, *2*(2), 75–99. https://doi.org/10.1080/15531180801958188

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Data mining concepts and techniques* (3rd ed.). https://doi.org/10.1109/ICMIRA.2013.45

Henderson, J., Wilson, A. M., Webb, T., McCullum, D., Meyer, S. B.,

Coveney, J., & Ward, P. R. (2017). The role of social media in communication about food risks. *British Food Journal*, *119*(3), 453–467. https://doi.org/10.1108/BFJ-07-2015-0272

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52. https://doi.org/10.1002/dir.10073

Jayasekara, P. K., & Abu, K. S. (2018). Text Mining of Highly Cited Publications in Data Mining. *IEEE 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, ETTLIS 2018*. https://doi.org/10.1109/ETTLIS.2018.8485261

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

Kaplan, J. J., Haudek, K. C., Ha, M., Rogness, N., & Fisher, D. G. (2014). Using lexical analysis software to assess student writing in statistics. *Technology Innovations in Statistics Education*, *8*(1), 107–132.

Kok, S. T., Wei, W. K., Goh, W., Yee, A., Chong, L., Weisfeld-Spolter, S., … Rabjohn, N. (2014). Examining the antecedents of persuasive eWOM messages in social media. *Online Information Review Corporate Communications An International Journal Iss Internet Research*, *38*(3), 746–768. https://doi.org/10.1108/OIR-04-2014-0089

Lim, W. M. (2016). Social media in medical and health care: opportunities and challenges. *Marketing Intelligence & Planning*, *34*(7), 964–976. https://doi.org/10.1108/MIP-06-2015-0120

Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58*(13), 2078–2091. https://doi.org/10.1002/asi.20672

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & Mccann, R. M. (2003). Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Communication Yearbook*, *27*(1), 293–335. https://doi.org/10.1207/s15567419cy2701_10

Moro, S., & Rita, P. (2018). Brand strategies in social media in hospitality and tourism. *International Journal of Contemporary Hospitality Management*,

30(1), 343–364. https://doi.org/10.1108/IJCHM-07-2016-0340

Oxford Dictionaries. (2018). Definition of social media in English. Retrieved April 4, 2018, from English Oxford Living Dictionaries website: https://en.oxforddictionaries.com/definition/social_media

Pantano, E., & Di Pietro, L. (2013). From e-tourism to f-tourism: emerging issues from negative tourists' online reviews. *Journal of Hospitality and Tourism Technology*, *4*(3), 211–227. https://doi.org/10.1108/JHTT-02-2013-0005

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463–479. https://doi.org/10.1509/jmr.12.0106

Udanor, C., Aneke, S., & Ogbuokiri, B. O. (2016). Determining social media impact on the politics of developing countries using social network analytics. *Program*, *50*(4), 481–507. https://doi.org/10.1108/PROG-02-2016-0011

Vraga, E. (2016). Party differences in political content on social media. *Online Information Review*, *40*(5), 595–609. https://doi.org/10.1108/OIR-10-2015-0345

Webopedia. (2018). UGC User-Generated Content. Retrieved April 4, 2018, from https://www.webopedia.com/TERM/U/UGC.html

Wu, L., Morstatter, F., Hu, X., & Liu, H. (2016). Mining misinformation in social media. In M. T. Thai, W. Wu, & H. Xiong (Eds.), *Big Data in Complex and Social Networks* (pp. 125–152). Boca Raton: Chapman and Hall/CRC.

Wu, L., Morstatter, F., & Liu, H. (2016). Misinformation key terms, explained. Retrieved April 5, 2018, from KD nuggets website: https://www.kdnuggets.com/2016/08/misinformation-key-terms-explained.html

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management*, *31*(2), 179–188. https://doi.org/10.1016/j.tourman.2009.02.016

Zawacki-Richter, O., & Naidu, S. (2016). Mapping research trends from 35 years of publications in distance education. *Distance Education*, *37*(3), 245–269. https://doi.org/10.1080/01587919.2016.1185079