The Application of Text Mining and Analytics Studies: A Systematic Literature Review

Shahid Shayya, Shamshul Bahri, Noor Ismawati Jaafar, Ainin Sulaiman, Seuk Wai Phoong, and Wai Chung Yeong

***Abstract*—A large amount of textual data is generated everyday through information technology, especially by social media platforms such as social network sites and mobile instant messaging applications. To analyse these large amount of textual data, analysts often turn to techniques called text mining and text analytics. Unfortunately, studies using these techniques are often more occupied with developing a new or extended models rather than determining how the findings could benefit organizations or societies. This occupation with the techniques rather than how the techniques could benefit the organizations or societies at large may render these studies a "plaything for the data scientists" rather than a useful technique to enhance knowledge and improve practice. This study intends to remedy this imbalance by identifying studies that use text mining and analytics techniques to inform organizational and societal practices. To do so, we will employ a method called the systematic literature review (SLR). The technique contains explicit and systematic process that distinguishes it from the conventional literature review. Eventually, the study reveals the source of data of the selected studies, their application area and the parties that will benefit from their findings. Lastly, this study discusses how studies using text mining and analytics can provide benefits to the larger society.**

***Index Terms*—Text analytics, text mining, systematic literature review, application.**

I. Introduction

Human beings are generating more and more data each day. This huge amount of data come in the form of pictures, videos and text. It is estimated that 80 percent of today's data are in unstructured form, which are expressed in rich and ambiguous natural language [1]. The large amount of textual data is generated from social media technologies such as Facebook and Twitter, and through mobile instant messaging apps such as WhatsApp and Telegram. It is estimated that five hundred million tweets are sent each day while forty million of those are shared. Meanwhile, it is estimated that 4.3 billion Facebook messages are posted with 5.75 billion likes daily [2].

To analyze the large amount of entirely unstructured data, a specific approach called text mining/analytics is required. Text mining is the predecessor to text analytics. Text mining is a subset of data mining. [3] uses natural language processing, knowledge management, data mining, and machine learning techniques to process text documents [4]. Text analytics, while similar to text mining in terms of method, usually deal with a bigger amount of data to extract and generate useful non-trivial information and knowledge [5].

Despite the large number of studies using text mining and analytics techniques, the impact they have on people has been unexplored. There has been too much emphasis on the techniques and less on how people can benefit from the findings. As a result, it is unclear how the findings from studies employing text mining and analytics have benefited the larger community. Hence, this study aims to investigate the human element in text mining and analytics research. To achieve this aim, we will systematically review the relevant literature that have employed both techniques.

The study has several significances. First, it refocuses the study using text mining and analytics to the human element. For far too long, researchers using both techniques are too preoccupied with the latest methods and models. We believe studies that neglect the application of the techniques for the benefit of the general population have very limited impact. Second, this study will help researchers, especially from the social science stream, to identify the evolution of the studies using text mining and analytics. Eventually, the researchers will be able to chart future research using the two techniques.

II. Literature Review

A. Text Mining and Analytics

Text mining and analytics have been applied in a number of fields. Two of those fields are finance and biomedical [6], [7]. [6] for example, employed text mining techniques to conduct a systematic review of studies on market prediction while [7] reviewed the applications of text mining in psychiatry. Whilst both studies employed the text mining techniques, their source of data vary widely, and hence their findings. [6] sourced the data for their study from online resources and market data such as The Wall Street Journal, Financial Times, Reuters and Bloomberg. Meanwhile, [7] sourced the data for their study from online databases such as CINAHL, Medline, EMBASE, PsycINFO and Cochrane.

The findings of both studies have some similarities. The most important similarity is the contribution of the studies to their respective fields. [6] helped structure the emerging field

Manuscript received May 19, 2018; revised August 17, 2018. This work was supported fully by the Malaysia Digital Economy Corporation (MDEC) under Grant PV003 - 2017 and conducted together with the Berkshire Media Private Ltd.

Muhammad Shahid Shayya is with the Berkshire Media Private Ltd, First Avenue Bandar Utama, 47800 Petaling Jaya, Selangor, Malaysia (e-mail: shahid@berkshiremedia.com.my).

Shamshul Bahri, Noor Ismawati Jaafar, Sulaiman Ainin, Seuk Wai Phoong, and Wai Chung Yeong are with the Department of Operations and MIS, Faculty of Business and Accountancy, University of Malaya, 50603 Kuala Lumpur, Malaysia (e-mail: esbi@um.edu.my, isma_jaafar@um.edu.my, ainins@um.edu.my, phoongsw@um.edu.my, yeongwc@um.edu.my).

DOI: 10.7763/IJCTE.2018.V10.1210

of market prediction and suggest areas of special significance which warrant further research. On the other hand, [7] identified how text mining can contribute to complex research tasks in psychiatry. They further discussed the benefits, limits and further application of the technique in the field.

III. Methodology

The main goal of this paper is to develop a deep understanding of the various text mining and text analytics studies that focussed its findings on organizational or societal applications. The technique used in the study is the systematic literature review. A review earns the adjective systematic if it is based on a clearly formulated question, identifies relevant studies, appraises their quality and summarizes the evidence by use of explicit methodology. It is the explicit and systematic approach that distinguishes systematic reviews from traditional reviews and commentaries [8]. The systematic literature review in this paper went through five (5) steps which we will describe in detail:

A. Framing Questions for a Review

In this study, we aimed to identify how research using text mining and analytics has informed human practices. We excluded the studies that focused on extending existing text mining and analytics models or developing new ones. Hence, the questions that we employed to frame the study were: (1) what are the studies that employed text mining and analytics but focused its findings on human application? And (2) how do these studies inform human practices? No changes or modifications were made to the questions throughout the study.

B. Identifying Relevant Work

The Web of Science was our sole database for the search of relevant articles. There were several reasons for our decision to use this single database. First, the strict selection of journals by the database convinced us that it is more authoritative than other academic databases. Second, it indexes the leading journals in the field of management, computer sciences, and information systems. These fields are closely related with the topics of this study. Third, the full article for most of the journals indexed in Web of Science are available through our university's library. It is often not the case for journals indexed by other entities. We started the identification of relevant work by using the keywords "text" and "mining" and "text' and "analytics". We then narrowed down the search using several criteria. First, we limited our search to articles that have those words in their titles. We believed that articles with those words in their titles will have a greater focus on the techniques compared to those which are not. Second, we limited our search to academic journals in the English Language because academic journals in the language represent the largest collection in the Web of Science. Third, we narrowed down the search to articles that were published between 2011 and 2016. We opined that the more recent articles will be more representative of the state-of-the-art research in the field. After using the three criteria, our search managed to find 362 articles that have the words "text" and "mining" and 122 articles that have the words "text" and "analytics" in their titles. We further narrowed down the number of articles by including those publications that are categorized as "articles" in the Web of Science and omitting the rest such as conference papers and books. We believe that academic articles present a more credible work in the field because they have gone through more intensive scrutiny by peers in the same discipline. The publications in the other categories often did not enjoy such scrutiny before they were published. As a result, we managed to reduce the number of articles on text mining to 411 and text analytics to 60.

A quick review of the articles that we had found at this stage revealed that there were still a large number of studies that focused on model development and extension, as opposed to the application of the findings. Therefore, we further narrowed down the scope of the search by identifying the relevant fields of study. We identified three relevant fields from the numerous fields of study offered by the Web of Science. The three disciplines are (1) computer science and information systems, (2) information systems and library sciences, and (3) social science: interdisciplinary. The fields of information systems and library sciences were chosen as they were the nearest fields of study to computer science that focussed on the adoption, development, and impact of information technology. Consequently, the studies would focus more on the human application side of the technology rather than the technology itself. This action reduced the number of articles on text mining to 69 while text analytics were reduced to 16.

C. Assessing the Quality of Studies

The next step of the study was to further distinguish studies that focused on the application rather than the development of model and application or extension of existing models. To achieve this aim, we went through the abstracts of the articles shortlisted. In the abstract, we searched for evidences that the study focused on the application of text mining and analytics in human practices. Inadvertently, we also searched for the opposite evidence; the studies that focused on method/model development or extension. This kind of study often proposes a model or method that it claims is superior to other models or methods. For example, [9] developed an ontology enrichment solution called ProMine that they claim to be superior in automatically identifying domain specific knowledge elements and automatic categorization of these extracted knowledge elements. We also searched for studies that focused on the development of applications or systems. For example, [10] developed BioTeks to support problem solving in life-sciences through text mining. Studies that fit both criteria were excluded from further analysis.

D. Summarizing the Evidence

Table I summarizes the articles on text mining and analytics that have been shortlisted for this study. The table also specifies the sources of the data, their application area, and the parties affected by the study. In this section, we will discuss the findings in detail.

Our analysis found that there are more articles on the topic of text mining than text analytics. In addition, we found that studies on text analytics only began to appear in 2012 while studies on text mining can be found since early 2000.

However, the number of articles on text analytics are steadily increasing for the last five years. This increase however does not affect the studies that are using text mining. Instead, there are still a large number of studies that use text mining as the method of analysis compared to text analytics.

We have also identified three main sources of data from the shortlisted studies. The first source of data was social media. From this source of data, we identified two of the most popular platforms used for data mining and text analytics: Facebook and Twitter. The second source of data came from organizations' internal databases. This source of data was more prevalent among healthcare organizations and business corporations. Meanwhile, the third source of data was full-text articles. This source of data represents academic studies that have been recorded in the form of journal articles. We found that this source of data for text mining and analytics was more popular among researchers.

In terms of application areas, we found them to be diverse. Although the highest number of application was found to be on business strategy, it was just slightly higher than the rest.

Some studies have been conducted in healthcare and information technology using text mining and analytics. For example, [11] used text mining techniques to gain insight from patient experience in the emergency department while [12] determined popular online shopping firms' Facebook patterns in Turkey.

We also found that studies using text mining and analytics have benefited a large number of people. On the one hand, the studies have benefited large business corporations as well as healthcare providers. For example, [13] analyzed technological proximities among patents application to support the decision making of merger and acquisition while in other example, [14] leveraged advanced text-mining techniques to identify self-disclosing health information on web forums. On the other hand, studies using text mining analytics have also benefited governments and also academic researchers. For example, [15] analyzed how citizens learn through social media while [16] analyzed 472 articles to study the scholarly development of the enterprise level IT innovation adoption literature.

TABLE I: SUMMARY OF ANALYSIS

Area of Studies	Source of Article	Source of Data	Application Area	Affected Parties
Text Mining	[17]	Interviews	Disaster management	Survivors and non-survivors; public safety agencies
	[18]	Twitter	Library	Libraries and their patrons
	[12]	Facebook	Information technology (E-commerce)	Online retailers and shoppers
	[19]	Full-text articles	Business strategy	Corporations
	[20]	Full text article	Information technology (E-commerce)	Researchers
	[21]	Full text articles	Healthcare (Biomedical)	Researchers
	[22]	Historical documents	Historical research	Historians
	[11]	Internal database	Healthcare	Patients and healthcare providers
	[23]	Patents records	Chemistry	Patent offices, patent applicants, and researchers.
	[24]	Job adverts	Information technology	Job seekers and companies.
	[14]	Web forums	Healthcare	Healthcare providers
	[13]	Patents records	Business strategy	Corporations
	[25]	Full-text articles	Healthcare (bio-informatics)	Researchers
	[26]	Facebook & Twitter	Business strategy	Corporations (fast-food companies, esp. pizza)
	[27]	Internal databases	Awards determination	Research funding agencies
	[28]	Internal databases	Business strategy	Corporations (airline companies)
Text Analytics	[5]	Full-text articles	Knowledge management	Researchers
	[15]	Social media	Government	Governments and citizens
	[29]	Internal databases	Business strategy	Corporations
	[30]	Internal databases	Healthcare	Healthcare providers and patients
	[16]	Full-text articles	Information technology (innovation & adoption)	Researchers

E. Interpreting the Findings

Although the terms text mining and text analytics have slightly difference meaning, they will continue to be used interchangeably in the near future. We also foresee the number of studies using text mining will still outnumber text analytics in the future. However, the gap between the two techniques will continue to shrink until a stage where the number of studies for both text mining and analytics will be equal.

This study dispels the myth that text mining and analytics techniques are useful only in the realm of social media. Text mining and analytics are equally useful in analysing organizations' internal databases. We also believe that more effort should be put into using text mining and analytics techniques to analyse the internal database. Organizations often collect huge amount of information from their customers, suppliers and even the staff. Unfortunately, many of these data are left stored in the servers. Organizations should be aware of the techniques that have the potential to provide insights into what is happening among their customers, suppliers and staff.

Text mining and analytics are useful in various areas outside the business context. It has been proven to be equally useful in the contexts of healthcare and information technology. Consequently, the use of the two techniques have benefited a large number of parties such as business corporations, healthcare providers, and academic researchers. We foresee text mining and analytics techniques will be employed in a wider context in the near future.

The techniques have already been deployed in disaster management [17] and historical research [22]. The expanding use of the techniques will only benefit a larger group of people in various communities.

IV. Conclusion

Despite the huge promises text mining and analytics hold, we are only scratching the surface of its potential. Organizations, especially, have not really leveraged the power of text mining and analytics to provide insights into their operational efficiency and strategy's effectiveness. However, things are beginning to change. More parties will be employing the two techniques, especially data analytics. Eventually, a larger group of people will benefit from their deployment.

Acknowledgment

We would like to thank the Malaysia Digital Economy Cooperation (MDEC) for its generous grant (Grant Number PV003-2017) that has enabled academics from the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya, and the staff of Berkshire Media Pt Ltd, Malaysia, to collaborate on this challenging and exciting research project.

References

[1] S. Debortoli, O. Müller, I. A. Junglas, and J. Brocke, "Text mining for information systems researchers: An annotated topic modeling tutorial," *Communications of the AIS*, vol. 39, no. 7, 2016.

[2] J. Schultz. (2016). How much data is created on the internet each day? [Online]. Available: https://blog.microfocus.com/howmuch-data-is-created-on-the-internet-each-day/

[3] A. Kaushik and S. Naithani, "A comprehensive study of text mining approach," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 16, no. 2, p. 69, 2016.

[4] F. Lucini, S. Fogliatto, G. J. C. Silveira, J. L. Neyeloff, M. J. Anzanello, R. S. Kuchenbecker, and B. D. Schaan, "Text mining approach to predict hospital admissions using early medical records from the emergency department," *International Journal of Medical Informatics*, vol. 100, pp. 1-8, 2017.

[5] Z. Khan and T. Vorley, "Big data text analytics: An enabler of knowledge management," *Journal of Knowledge Management*, vol. 21, no. 1, pp. 18-34, 2017.

[6] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, pp. 7653-7670, 2014.

[7] A. Abbe, C. Grouin, P. Zweigenbaum, and B. Falissard, "Text mining applications in psychiatry: A systematic literature review," *International Journal of Methods in Psychiatric Research*, vol. 25, no. 2, pp. 86-100, 2016.

[8] K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes, "Five steps to conducting a systematic review*," Journal of the Royal Society of Medicine*, vol. 96, no. 3, pp. 118-121, 2003.

[9] S. Gillani and A. Ko, "Incremental ontology population and enrichment through semantic-based text mining: an application for it audit domain," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 11, no. 3, pp. 44-66, 2015.

[10] R. Mack, S. Mukherjea, A. Soffer, N. Uramoto, E. Brown, A. Coden, and H. Matsuzawa, "Text analytics for life science using the unstructured information management architecture," *IBM Systems Journal*, vol. 43, no. 3, pp. 490-515, 2004.

[11] P. Varanasi and M. Tanniru, "Seeking intelligence from patient experience using text mining: Analysis of emergency department data," *Information Systems Management*, vol. 32, no. 3, pp. 220-228, 2015.

[12] E. Kahya-Özyirmidokuz, "Analyzing unstructured Facebook social network data through web text mining: A study of online shopping firms in Turkey," *Information Development*, vol. 32, no. 1, pp. 70-80, 2016.

[13] H. Park, J. Yoon, and K. Kim, "Identification and evaluation of corporations for merger and acquisition strategies using patent information and text mining," *Scientometrics*, vol. 97, no. 3, pp. 883-909, 2013.

[14] Y. Ku, C. Chiu, Y. Zhang, H. Chen, and H. Su, "Text mining self-disclosing health information for public health service," *Journal of the Association for Information Science and Technology*, vol. 65, no. 5, 928-947, 2014.

[15] C. G. Reddick, A. T. Chatfield, and A. Ojo, "A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government Facebook use," *Government Information Quarterly*, vol. 34, no. 1, pp. 110-125, 2017.

[16] R. C. Basole, C. D. Seuss, and W. B. Rouse, "IT innovation adoption by enterprises: Knowledge discovery through text analytics," *Decision Support Systems*, vol. 54, no. 2, pp. 1044-1054, 2013.

[17] N. Y. Yun and S. W. Lee, "Analysis of effectiveness of tsunami evacuation principles in the 2011 Great East Japan tsunami by using text mining," *Multimedia Tools and Applications*, vol. 75, no. 20, pp. 12955-12966, 2016.

[18] S. M. Al-Daihani and A. Abrahams, "A text mining analysis of academic libraries' tweets," *The Journal of Academic Librarianship*, vol. 42, no. 2, pp. 135-143, 2016.

[19] K. W. McCain, "Mining full text journal articles to assess obliteration by incorporation: Herbert A. Simon's concepts of bounded rationality and satisficing in economics, management, and psychology," *Journal of the Association for Information Science and Technology*, vol. 66, no. 11, pp. 2187-2201, 2015.

[20] B. N. Yan, T. S. Lee, and T. P. Lee, "Analysis of research papers on E-commerce (2000-2013): Based on a text mining approach," *Scientometrics*, vol. 105, no. 1, pp. 403-417, 2015.

[21] X. Zhai, Z. Li, K. Gao, Y. Huang, L. Lin, and L. Wang, "Research status and trend analysis of global biomedical text mining studies in recent 10 years," *Scientometrics*, vol. 105, no. 1, pp. 509-523, 2015.

[22] U. Hinrichs, B. Alex, J. Clifford, A. Watson, A. Quigley, E. Klein, and C. M. Coates, "Trading consequences: A case study of combining text mining and visualization to facilitate document exploration," *Digital Scholarship in the Humanities*, vol. 30, no. 1,pp. i50-i75, 2015.

[23] Y. Ju and S. Y. Sohn, "Identifying patterns in rare earth element patents based on text and data mining," *Scientometrics*, vol. 102, no. 1, pp. 389-410, 2015.

[24] S. Debortoli, O. Müller, and J. Brocke, "Comparing business intelligence and big data skills," *Business & Information Systems Engineering*, vol. 6, no. 5, pp. 289-300, 2014.

[25] M. Song and S. Y. Kim, "Detecting the knowledge structure of bioinformatics by mining full-text collections," *Scientometrics*, vol. 96, no. 1, pp. 183-201, 2013.

[26] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464-472, 2013.

[27] D. McArthur and H. Crompton, "Understanding public access cyberlearning projects using text mining and topic analysis," *Journal of the Association for Information Science and Technology*, vol. 63, no. 11, pp. 2146-2152, 2012.

[28] R. S. Rodrigues, P. P. Balestrassi, A. P. Paiva, A. GarciaDiaz, and F. J. Pontes, "Aircraft interior failure pattern recognition utilizing text mining and neural networks," *Journal of Intelligent Information Systems*, vol. 38, no. 3, pp. 741-766, 2012.

[29] O. Müller, S. Debortoli, I. Junglas, and J. Brocke, "Using text analytics to derive customer service management benefits from unstructured data," *MIS Quarterly Executive*, vol. 15, no. 4, 2016.

[30] S. Boytcheva, G. Angelova, Z. Angelov, and D. Tcharaktchiev, "Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care," *Cybernetics and Information Technologies*, vol. 15, no. 4, pp. 58-77, 2015.

Muhammad Shahid Shayaa is the founder and the CEO of Berkshire Media. Shahid is instrumental in transforming the digital media industry through rigorous data driven approaches. He has accumulated more than 13 years of experience working in the corporate sector in various business transformation projects in Accenture, Employees Provident Fund, Malaysia Airlines, and the Government of Malaysia's Performance Management Delivery Unit. He is now advises the top leaderships on strategic communications for publicly traded companies and the public sector.

Noor Ismawati Jaafar is an associate professor and deputy dean for research & development in the Faculty of Business and Accountancy, University of Malaya. Her research interests include accounting information systems, information technology management, information technology governance and social media. She has published papers in Information & Management, Government Information Quarterly, Computers in Human Behavior, Telematics and Informatics, Behavior & Information Technology, Cyberpsychology, Behavior and Social Networking, Information Development, International Journal of Mobile Communications and so on.

Sulaiman Ainin is a professor in the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya. Her research interest includes management information systems, technology diffusion, e-commerce and green information technology. She has published papers in Information & Management, Computers in Human Behavior, Telematics and Informatics, Behavior & Information Technology, American Journal of Scientific Research, Management Decisions, International Journal of Mobile Communications, Government Information Quarterly and so on.

Seuk Wai Phoong is a senior lecturer in the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya. Her research areas are in econometrics modeling, time series economics, and statistical modeling and spatial analysis. She has published articles in International Journal of Advanced and Applied Sciences and International Journal Computing Science and Mathematics.

Wai Chung Yeong is a senior lecturer in the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya. His research areas are in optimization, total quality management, operational effectiveness and economic designs. He has published in many ISI-indexed journals including European Journal of Operational Research, Quality Technology and Quantitative Management, Computers & Industrial Engineering, Journal of Quality Technology and Quality Engineering.

Shamshul Bahri is a senior lecturer and head of the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya. His areas of research are in information, computer and communication technology (ICT) and information systems (public health informatics). He is a prolific qualitative researcher and has published articles in leading IS journals such as Information & Management and Information Systems Journal.