Doctoral Education in Context

Perspectives from Africa

Edited by
Jan Botha, Liezel Frick, Nompilo Tshuma

AFRICAN
MINDS

Published in 2024 by African Minds
4 Eccleston Place, Somerset West, 7130
Cape Town, South Africa
info@africanminds.org.za
www.africanminds.org.za

All contents of this document, unless specified otherwise, are licensed under a Creative Commons Attribution 4.0 International License.



ISBNs:
978-1-0672535-3-0 Print
978-1-0672535-4-7 e-book
978-1-0672535-5-4 e-pub

Copies of this book are available for free download at www.africanminds.org.za

ORDERS
For orders from Africa:
African Minds
Email: info@africanminds.org.za

For orders from outside Africa:
African Books Collective
PO Box 721, Oxford OX1 9EN, UK
Email: orders@africanbookscollective.com

CHAPTER 1

Bibliometric analysis of doctoral supervision literature using topic modelling

Rudolph Oosthuizen

South Africa has a population of 62 million. The country has 26 public universities with a total enrolment of 1,1 million students. Doctoral enrolments in South Africa have increased from 9 994 in 2008 to 23 588 in 2020 (CHE Vitalstats). Between 2000 and 2018, a total of 32 025 doctoral students graduated at South African universities. Doctoral graduates increased from 972 in 2000 to 3 339 in 2019 (SciSTIP Tracer Study).

The **University of Pretoria** (UP) became an independent university in 1930, tracing its origins to the Normal College (est. 1902) and the Transvaal Technical Institute (est. 1904) which merged to form the Transvaal University College (1908). UP has 2 217 academic staff members working in nine faculties and a business school, with a total student enrolment of 53 900, which makes it one of the largest contact-education universities in Africa. In 2021 UP had 2 537 doctoral candidates.

1. Introduction

The doctorate's importance, worldwide and in Africa, has increased over the past few decades. Generally, the aim is to provide researchers with high-level skills required by the knowledge economy, placing a renewed focus on the doctoral supervisor's contribution and role (Cyranoski et al., 2011).

The South African Qualification Standard for Doctoral Degrees defines the purpose of a doctoral degree as to "develop the highest level of holistic and systematic understanding of scholarship in, and stewardship of, a field of study

through an original contribution that advances the frontiers of knowledge" (CHE, 2018, p. 11). Doctoral studies need to display mastery in and develop appropriate research methods and skills to pursue knowledge in an inter-disciplinary and professional manner. Doctoral supervision plays a significant role in achieving these aims. Adequate doctoral supervision relies on the supervisor's and the student's skills, characteristics, and traits (CHE, 2018).

Strengthening the practice and knowledge of doctoral supervision requires research reported in scholarly articles. This chapter aims to extract and analyse the research topics in peer-reviewed articles on doctorate supervision. These topics were examined for structure and trend aspects to understand the current state of research in this field and to identify future opportunities and needs. The first step is to discuss using bibliometrics to support the planning and conduct of the research. Natural Language Processing (NLP), supported by data mining and topic modelling, is presented as a method that can be used to analyse bibliometric data. The next section explains the process of executing doctoral supervision literature topic modelling. Processing and analysing the captured data will provide insight into doctoral supervision research by discussing the topics' relative importance and temporal trends. This chapter concludes by discussing the topic modelling results to highlight the key findings.

2. Bibliometrics and Natural Language Processing

Scientists and researchers codify their findings in articles for journals, conferences or scholarly books. Because these articles are peer-reviewed with expert-based judgment, these building blocks of science and knowledge are validated (Hood & Wilson, 2001; Keathley-Herring et al., 2015). Bibliometric analysis is a systematic analysis method to analyse published research. Visualisation of the bibliometric outputs helps identify publication patterns and trends in a research field (Jie et al. 2014, Jiang et al., 2016). Publication and citation history trends may help researchers evaluate a field's prominent authors, growth, performance, and maturity (Aria & Cuccurullo, 2017; Kalantari et al., 2017). The most popular bibliometric indicators include citations, number of articles published, author statistics, keyword trend analysis, relational indicators (co-words, co-publication and co-citations), or research topics (Jia et al., 2018). Topic modelling can extract the core topics (concepts or themes) from a body of literature and calculate the trends (Lamba & Madhusudhan, 2019).

Traditionally, the manual bibliometric analysis process assigned articles to a predetermined topic list proposed by experts in the field. The number of articles available from journals and published conference proceedings has

increased exponentially in most fields. Reading and sorting all of these articles into topics is time-consuming, challenging, and risks being biased. The manual analysis may also miss the hidden, latent, or emerging topics in a large text corpus. Automated topic modelling, implemented through machine-learning-based methods, will improve this process and eliminate or reduce these risks (Lee & Kang, 2018). This chapter focuses on automated topic modelling to identify the research trends in doctoral supervision literature.

Software analytics tend to struggle with processing unstructured text and free-format text. NLP converts unstructured text into numerical values possible for machine-learning algorithms to process, among other uses. NLP can extract relationships, key phrases, and patterns in a text corpus. Therefore, NLP is mainly applied to categorise and cluster text for document classification and summarisation to enable information extraction. Other uses include speech recognition and language translation (Banu & Chitra, 2015; Agrawal et al., 2018).

Machine-learning text-mining approaches have become more usable in recent times as the software, information, and computing power become increasingly accessible for researchers. Topic modelling provides an unsupervised text classification tool with quantitative statistical algorithms to extract semantic information from text. The computational algorithms automatically process a large text corpus to discover latent topics. Topic modelling assumes that words in a document constitute a set of intentional and latent topics. Topic modelling is independent of prior knowledge of the corpus (Jiang et al., 2016; Agrawal et al., 2018).

Latent Dirichlet allocation (LDA) is a popular topic modelling implementation algorithm. LDA identifies major thematic clusters from an extensive corpus of text documents to allocate each topic to each document. However, LDA requires input from the researcher on the number of issues for extraction. The algorithm extracts a distribution of keywords to define the topics. Despite being an unsupervised method, expert and domain knowledge are still required to name and describe the extracted topics (Tong & Zhang, 2016; Suominen & Toivanen, 2016; Hecking & Leydesdorff, 2019; Maier et al., 2018).

Despite some challenges, such as being non-deterministic and dependent on input parameters, topic modelling is more comprehensive and faster than manual methods. Topic modelling outputs enable researchers to quickly process a large corpus of articles for a more in-depth analysis of key elements within the literature. The automated process also provides information already in digital format for easier and more effective post-processing and reporting (Maier et al., 2018; Asmussen & Møller, 2019).

3. Method

The author applied the method discussed in this chapter in other fields, such as Systems Engineering, Technology Management and Sustainable Development (Oosthuizen & Pretorius, 2020).

3.1 Data capture

The research process used to process and analyse the titles and abstracts from articles on doctoral supervision is depicted in Figure 1. This research process uses Python-based algorithms and functions with the support of Microsoft Excel. A prerequisite for an automated topic modelling process is that the bibliometric data on the relevant published research is accessible in digital format. The articles for this analysis were extracted from the Scopus abstract and citation database using the search terms **"(doctoral OR doctorate OR PhD OR Ph.D) AND supervision**" over the period 1975-2018. The data captured in each article included the following:

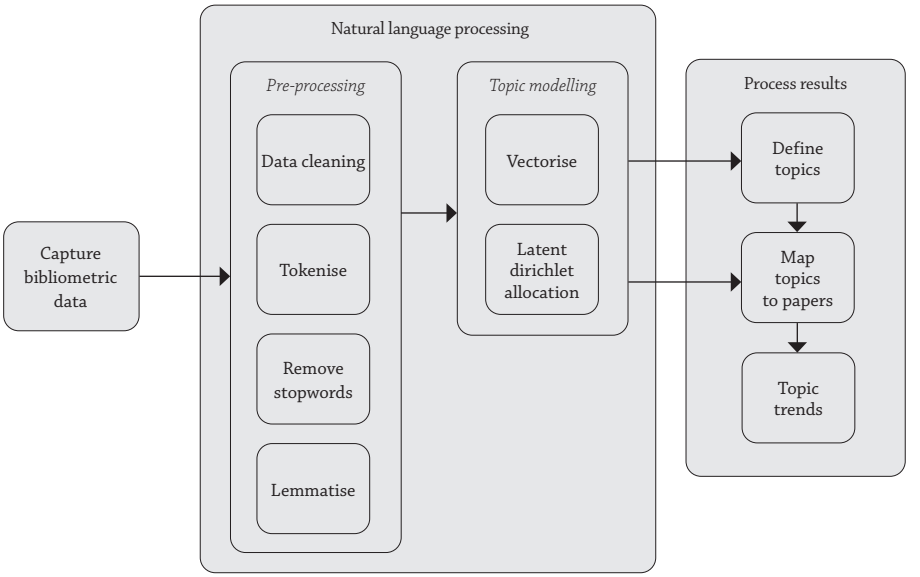
1. Year of publication.
2. Authors.
3. Title.
4. Abstract.

3.2 Pre-processing

The SpaCy library in Python implemented the NLP algorithms. The raw captured text required pre-processing to prepare it for analysis. The first step removed articles not relevant to the field of doctoral supervision or not containing all the required information (e.g. authors, title, or abstract). Next, the titles and abstracts were combined to increase the text data's size to be processed. Titles of articles contain the most relevant words for the article, while the abstract summarises the problem and the research results.

Data cleaning also included the removal of publisher-related text (e.g. "Wiley & Sons", "Elsevier", "all rights reserved", etc.) and fixing spelling errors. Next, the process removed punctuation (i.e. periods, commas, exclamation points, ampersands, etc.), whitespaces, letters, numbers, and special characters that could affect the NLP algorithms. Although punctuation adds meaning to the text for a human reader, it is undesirable and uninformative for the NLP algorithm.

Figure 1: Bibliometric analysis process



Tokenisation uses the spaces in the text to extract the linguistic units that represent the building blocks for sentences or paragraphs (Patel & Soni, 2012; Lin et al., 2016). Next, stop words (common words in a language that do not add real meaning to the text, such as "and", "the", "if", "a", etc.) are removed to reduce the dimensionality of the text (Maier et al., 2018). Lemmatisation also reduces the text corpus's dimensionality by deriving the base word by calculating each term's lemma. Python's SpaCy library was used to implement these steps.

Finally, generic words not contributing to identifying topics were also removed. The output of several test runs of the topic modelling software were manually investigated to identify words not contributing to the identification of the topics. These include common words due to the academic nature of the published articles, such as "represent", "aim", "paper", "example", and "define". Cleaning and preparing the text data for topic modelling is one of the most important steps to ensure quality topics are extracted.

3.3 Topic modelling

In this research, topic modelling is the key step in the process of bibliographic analysis. The Scikit-learn library in Python is used to vectorise the text and

perform topic modelling with an LDA algorithm. The outputs of the LDA application provide the following:

1. The required number of topics in the corpus and their describing terms (words and phrases).
2. The relative importance (probability) of each term for the topic in the corpus.
3. The probability of each topic to be represented per document.
4. Perplexity value of the whole topic model.

The random number generator initiating the LDA algorithm training causes each software run to produce an output set of topics with some variation between the different runs. Therefore, selecting a suitable set of LDA parameters will improve the LDA model's stability (Hecking & Leydesdorff, 2019). Perplexity provides a measure to determine the topic model's statistical quality or goodness of fit (Blei et al., 2003). The lower the perplexity score, the better the model. Calculating the perplexity scores for different combination sets of parameter values for the model determines an optimum set of parameters (Hagen, 2018). The selected parameters must consider interpretability, replicability, external validity, and internal coherence between documents and topics (Isoaho et al., 2019).

3.4 Process results

Since the LDA algorithm only clusters documents by their topics, a manual analysis is still required to interpret and name each topic without indicating it. One way to assist interpretation is by using word clouds. The list of weighted terms is presented in a word cloud to assess the most important terms at a glance (Bashri & Kusumaningrum, 2017). Usually, a group of experts will be consulted, but the author performed it himself in the case of this chapter.

Calculating the topic popularity and trends utilises the probability of topic allocation per article. Multiple topics may be allocated per article since there is an overlap between topics. Therefore, using the most probable topic assigned per article is not the best solution. Also, because the total number of articles published per year is not constant, the data were transformed by dividing the total articles per topic by the total number of published documents. This provides the relative importance of topics in the published research on doctoral supervision.

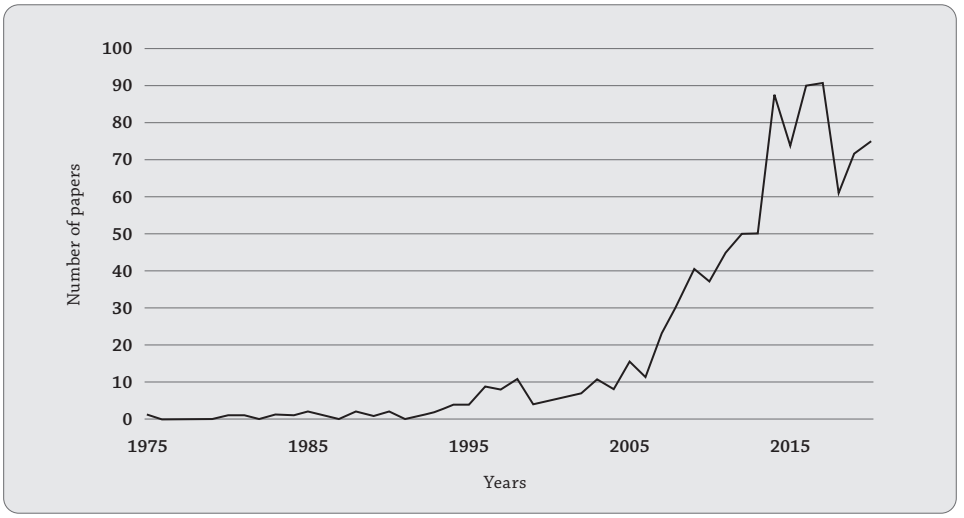
4. Results

4.1 Pre-processing and parameter selection

At the start of the process, 1 467 journal articles, book chapters, and published conference proceedings were captured from the Scopus database with a Python algorithm using the search terms discussed above. Although the Scopus database may be limited regarding publications of doctoral supervision research conducted in Africa, it still provides a starting point for investigating the research topics in the field. The first cleaning exercise reduced the number of articles to 949 by manually removing articles not relevant to the field of doctoral supervision, as judged by the author. Only English articles were included in the sample to facilitate the NLP algorithms. Other articles removed include errata, corrections, editorials, and articles with incomplete data. Figure 2 shows a noticeable increase in articles published on doctoral supervision since the early 2000s, peaking around 2017. The graph in Figure 2 seems to agree with the findings of Cyranoski et al., (2011).

In addition to the pre-processing steps presented in section 3.2, the terms "doctoral", "supervision", "phd", and "supervise" were also removed before extracting the topics. Since these were the document search terms, they tend to be some of the most numerous words in the data sample. This project aims to find research topics published in the field of doctoral supervision. An extracted topic named "Doctoral Supervision" will add no value to the objective of this

Figure 2: History of articles on doctoral supervision

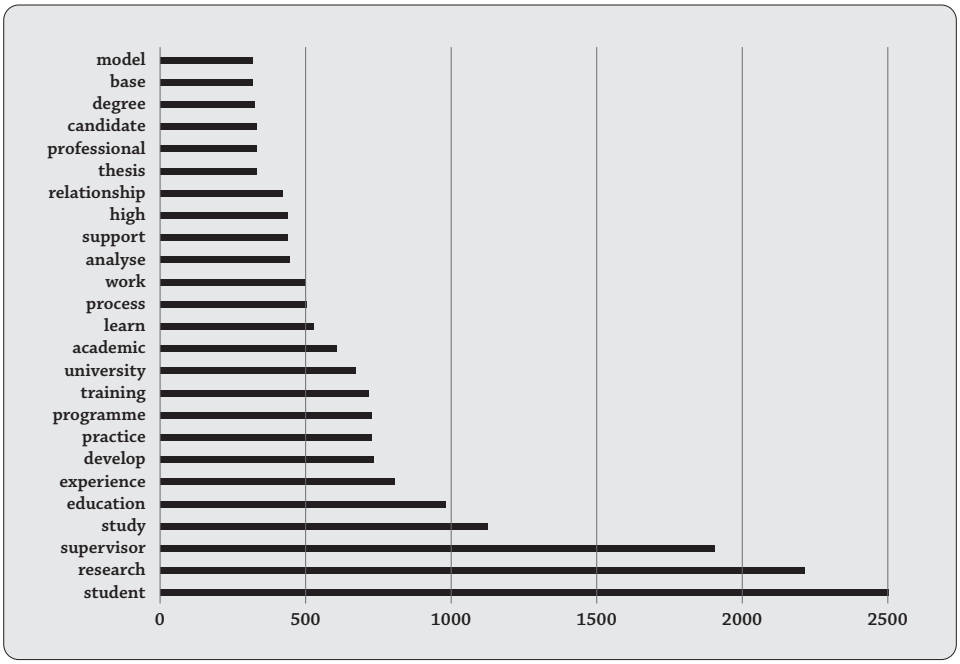


analysis. Therefore, removing these terms will ensure that the extracted topics are "within" the field of doctoral supervision. Figure 3 shows the word counts of the top 25 terms remaining for topic modelling processing.

The three main parameters that affect the topic extraction process are the minimum and maximum document frequency of the words in the processed text and the number of required topics for extraction. The minimum and maximum document frequency ensures that terms that occur in too few or too many documents are ignored. These terms affect the size of the data available for processing and the number of topics that can be extracted. The number of documents and their size affect the selection of values for these parameters.

After comparing the perplexity scores, a minimum document frequency of 0.13 and a maximum document frequency of 0.95 were selected. Therefore, the terms have to occur in more than 13% and less than 95% of the text corpus articles. The number of topics was chosen as five (5) as it was assumed that research into "Doctoral Supervision" should not result in too many straightforward and easy-to-name topics.

Figure 3: Word counts of the top 25 terms to be processed



4.2 Doctoral supervision research topics

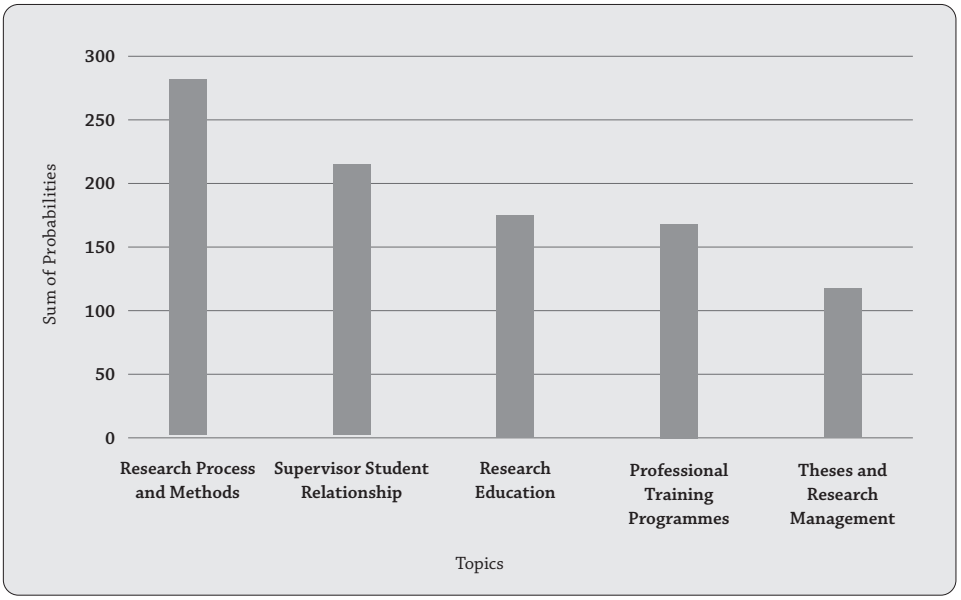
The topics extracted from the captured articles on doctoral supervision are listed, named and described in Table 1. The table presents the derived name and word cloud of each topic extracted from the extracted article's title and abstract.

The topic modelling process also provides the probability of each of the five topics allocated to each article in the text corpus. Summing up these ratios per article indicates the popularity of each topic over the publication history to provide a topic frequency, as seen in Figure 4. The "Research Process and Methods" topic is most prevalent in the processed articles. Another popular topic is about the relationship between supervisors and their doctoral students. The third and fourth most popular topics are about the students' research and professional skills training to equip doctoral students. Lastly, the least popular topic covers research and study process management.

Table 1: Topics with word clouds

Number	**1**	**2**	**3**
Name	Research process and methods	Supervisor–student relationship	Research education
Wordcloud			
Number	**4**	**5**	
Name	Professional training programmes	Theses and research management	
Wordcloud			

Figure 4: Topic frequency

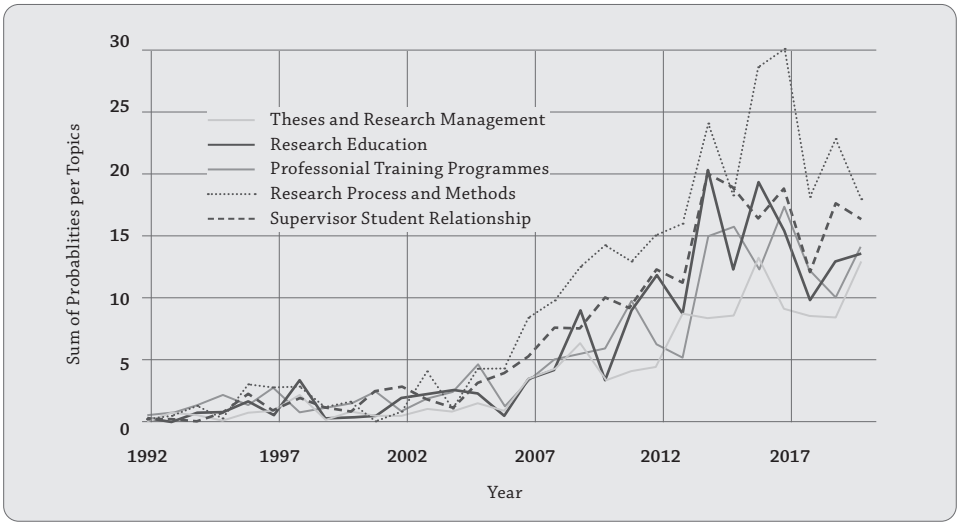


4.3 Research publication trends

The NLP and topic modelling process's data output provides a rich dataset for processing, depending on the focus question. This section generates trends for the extracted topics extracted from the articles on doctoral supervision to understand the current focus and identify future opportunities and needs for research. Figure 5 shows the trend per year of the five topics. The popularity is calculated by summing the allocation probability per article per topic per year. The data is only shown from 1992 to 2018 because the previous years contained only a few articles.

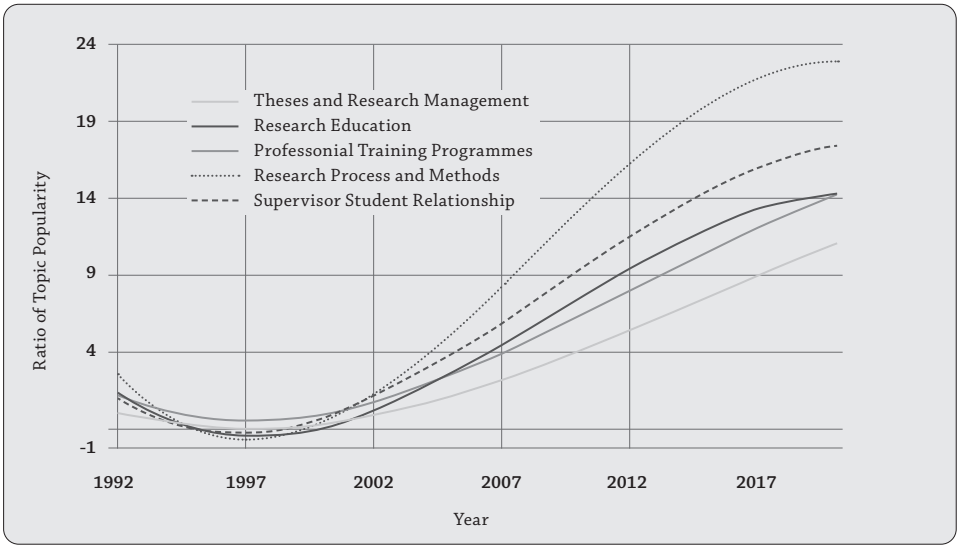
Since the graph in Figure 5 is not easy to interpret, smoothened trends are calculated using a fourth-order binomial regression due to the year-on-year variability. As seen in Figure 6, the trend lines for all the topics are on a steep, increasing trajectory. However, these trends are closely related to the annual increase in the total number of articles per year on doctoral students' supervision from Figure 2. This provides no information on the trends of the topics relative to each other.

Figure 5: Popularity per Topic



Since the number of articles per year is not constant, the data is transformed by dividing the topic allocation values per year by the number of articles for that year. This enables the calculation of the relative popularity between the topics per year. Figure 7 presents the transformed topic trends over the analysed publication years. However, the few articles during the early years resulted in

Figure 6: Smoothened topic trends



more widely disparate numbers. The later years, with a large number of articles, caused smoother graphs with subtle trends. However, this makes the identification of trends more difficult.

Despite being the most popular topic, "Research Process and Methods" has experienced a slight decline in the last decade. The topic of "Research Education" presents a similar trend. The "Professional Training Programmes" topic is experiencing a constant trend relative to the other topics. The two topics with an upward trend relative to the other topics are "Theses and Research Management" and "Supervisor–student Relationship".

5. Discussion

In this paper, the LDA process extracted a set of 5 topics from the titles and abstracts of 949 articles on the theme of doctoral supervision. Regardless of the limitations of automated unsupervised machine-learning- based topic modelling, it can still provide valuable support for doctoral education researchers when they do their literature studies. An in-depth analysis of the output topics could also support the development of a comprehensive list of research topics in the field. The five extracted topics can be grouped into two main themes, as seen in Table 2.

Table 2: Grouping of research topics

Educational or training topics	Supervisor roles
Research Education Professional Training Programmes Research Process and Methods	Supervisor–student Relationship Theses and Research Management

The extracted topics and the main themes from Table 2 also agree with the article by Bastalich (2017). In the article, she noted the importance of improving the supervision relationship within the context of political regulation, pedagogy and academic socialisation. The article also discusses the roles of research training and student and supervisor development.

Another interesting finding is the difference between "Research Education" and "Professional Training Programmes" topics. The data shows that both seem important for doctoral supervision as they exhibit similar trends. Doctoral supervision tends to be conceptual, a broad, knowledge-based educational, formative endeavour that requires the skills to perform the requisite activities. The topic of "Research Education" contains terms such as "research", "academic", "work", "quality", "knowledge", and "higher", while "Professional

Training Programmes" focus on "practice", "skill", "develop" and "experience." This shows that the two perspectives on doctoral supervision development can be treated equally and separately. "Education" seems to be more conceptual, comprehensive or formative (of the whole human being) in nature and "training" focuses on practical skills required during doctoral supervision (Bastalich, 2017).

Developing the required doctoral supervision skills requires a pedagogical approach. This is based on positive relationships supported by critical thinking skills. Students must also be exposed to scholarly research methods (Frick, 2019). These elements are present in the extracted topics. Botha and Mouton (2019) listed the primary task of the supervisor as the following:

1. Supervision of the research process through advice (relationship) and instruction (research education and training).
2. Coaching as part of the supervision relationship.
3. Project management.
4. Basic prerequisites for supervision (presence, commitment, and mutual conversation).

These tasks are compared to the extracted topics in Table 3. The table clearly shows that all the tasks can be related to one or more topics. This provides a form of validation for the extracted topics.

Table 3: Comparison of topics to supervisor tasks

No.	Topic	Supervision of the research process	Coaching	Project management	Basic prerequisites for supervision
1	Theses and Research Management			X	X
2	Research Education	X			
3	Professional Training Programmes	X			
4	Research Process and Methods	X			
5	Supervisor–student Relationship	X	X		X

6. Conclusion

NLP, implemented through unsupervised machine-learning methods, was applied to perform a bibliometric study on doctoral supervision research,

published from 1975 to 2018. The analysis's output presents valuable bibliometric information to support doctoral students' supervision and development research. It is important to note that the set of output topics resulted from unsupervised machine-learning, which may not be able to produce the perfect set of topics during each execution of the probabilistic algorithm. The selection of the parameters and the quality of the processed text data are significant factors in this process. However, the outputs are good enough to investigate the field of doctoral supervision and identify research opportunities within an extensive text dataset.

These topics can help researchers understand the current and historic research state and trends in doctoral supervision research. The output data provide rich information for further analyses to investigate specific questions about systems engineering. The topic model can be refined to give access to more, and detailed, topics in future work. These tools and methods may also be used to support literature reviews.

7. References

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology, 98*, 74–88.

Antons, D., Kleer, R., & Salge, T. O. (2016). Mapping the topic landscape of JPIM, 1984–2013: In search of hidden structures and development trajectories. *Journal of Product Innovation Management, 33*(6), 726–749.

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975.

Asmussen, C. B., & Møller, C. (2019). Smart literature review: A practical topic modelling approach to exploratory literature review. *Journal of Big Data, 6*(1), 93.

Banu, G. R., & Chitra, V. (2015). A survey of text mining concepts. *International Journal of Innovations in Engineering and Technology, 5*(2), 121–127.

Bashri, M. F., & Kusumaningrum, R. (2017). *Sentiment analysis using Latent Dirichlet Allocation and topic polarity wordcloud visualisation*. 5th International conference on information and communication technology (ICoIC7), Melaka, Malaysia 1–5. https://ieeexplore.ieee.org/document/8074651

Bastalich, W. (2017). Content and context in knowledge production: A critical review of doctoral supervision literature. *Studies in Higher Education, 42*(7), 1145–1157.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(January), 993–1022.

Botha, J. (2019). *The nature, purpose, standard and format of the doctoral degree*. Course material of module 2 of the DIES/CREST training course for supervisors of doctoral candidates at African universities. Stellenbosch University.

Botha, J., & Mouton, J. (2019). *The preparation phase: Selection, supervisor allocation, and supervising the doctoral proposal*. Course material of module 5 of the DIES/CREST Training course for supervisors of doctoral candidates at African universities. Stellenbosch University.

Council on Higher Education (CHE). (2018). *Qualification standard for doctoral degrees*. CHE.

Cyranoski, D., Gilbert, N., Ledford, H., Nayar, A., & Yahia, M. (2011). The PhD factory: The world is producing more PhDs than ever before. Is it time to stop? *Nature, 472*(7343), 276–279.

Eker, S., Rovenskaya, E., Langan, S., & Obersteiner, M. (2019). Model validation: A bibliometric analysis of the literature. *Environmental Modelling & Software, 117*, 43–54.

Frick, L. (2019). *Supervisory models and styles*. Course material of module 4 of the DIES/CREST Training course for Supervisors of doctoral candidates at African universities. Stellenbosch University.

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management, 54*(6), 1292–1307.

Hecking, T., & Leydesdorff, L. (2019). Can topic models be used in research evaluations? Reproducibility, validity, and reliability when compared with semantic maps. *Research Evaluation, 28*(3), 263–272.

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics, 52*(2), 291–314.

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2019). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal, 49*(1), 300–324.

Jia, Y., Wang, W., Liang, J., Liu, L., Chen, Z., Zhang, J., Chen, T., & Lei, J. (2018). Trends and characteristics of global medical informatics conferences from 2007 to 2017: A bibliometric comparison of conference publications from Chinese, American, European and the Global Conferences. *Computer Methods and Programs in Biomedicine, 166*, 19–32.

Jiang, H., Qiang, M., & Lin, P. (2016). A topic modeling based bibliometric exploration of hydropower research. *Renewable and Sustainable Energy Reviews, 57*, 226–237.

Jie, L., Xiaohong, G., Shifei, S., & Jovanovic, A. (2014). Bibliometric mapping of "International symposium on safety science and technology (1998-2012)". *Procedia Engineering, 84*, 70–79.

Kalantari, A., Kamsin, A., Kamaruddin, H. S., Ebrahim, N. A., Gani, A., Ebrahimi, A., & Shamshirband, S. (2017). A bibliometric approach to tracking big data research trends. *Journal of Big Data, 4*(1), 30.

Keathley, H., Bean, A., Chen, T., Vila, K., Ye, K., & Gonzalez-Aleu, F. (2015). Bibliometric analysis of author collaboration in engineering management research. In E.-H. Ng, S. Long & A. Squires (Eds.), *Proceedings of the international annual conference of the American society for engineering management* (pp. 679–689). American Society for Engineering Management.

Lamba, M., & Madhusudhan, M. 2019. Mapping of topics in DESIDOC Journal of Library and Information Technology, India: a study. Scientometrics 120 (2), 477-505.

Lee, H., & Kang, P. (2018). Identifying core topics in technology and innovation management studies: A topic model approach. *The Journal of Technology Transfer, 43*(5), 1291–1317.

Lin, J. R., Hu, Z. Z., Zhang, J. P., & Yu, F. Q. (2016). A natural-language-based approach to intelligent data retrieval and representation for cloud BIM. *Computer-Aided Civil and Infrastructure Engineering, 31*(1), 18–33.

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

Oosthuizen, R., & Pretorius, L. (2020, July 18–23). *A bibliometric method for analysis of systems engineering research* [Paper presentation]. 30th annual INCOSE international symposium, Cape Town, South Africa.

Patel, F. N., & Soni, N. R. (2012). Text mining: A brief survey. *International Journal of Advanced Computer Research, 2*(6), 243.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464–2476.

Tong, Z., & Zhang, H. 2016. A Text Mining Research Based on LDA Topic Modelling. Proceedings of the Sixth International Conference on Computer Science, Engineering and Information Technology (CCSEIT 2016), Vienna, Austria, May 21-22, 2016, pp.201-210.