Evaluating Publication Keywords in Computer Science Education Research - A Bibliometric NLP Approach

Abstract
This work demonstrated how evaluating publication keywords in the Computer Science Education Research (CSER) could bring conceptual and functional insights by combining the bibliometric approach and natural language processing (NLP). The collection of publication keywords represents the knowledge landscape of the research domain. Using proper keywords will improve publication visibility in research networks, contributing to the overall research impact. We gathered bibliometric data and the strategic directions in two CSER publication venues from 2015 to 2019 to capture the research foci of CSER and evaluate alignment between 1) selected keywords and publication venue mission statements; 2) keywords and abstract content. By applying the NLP techniques, our results revealed that the most prevalent research foci represented by the most commonly used CSER keywords were teaching & learning and broadening participation, which aligned with the corresponding strategic directions. However, our analysis also suggested a misalignment between keywords identified by authors and the topics presented in the abstracts. With our work, we hope to motivate scholars to carefully evaluate and select keywords for indexing publications so as to improve research topic relevancy and publication visibility for broader impacts.

1 Introduction

In the rapidly growing Computer Science Education Research (CSER) field, the collection of publication keywords represents a concise knowledge landscape of the research domain. Understanding this landscape helps stakeholders contextualize existing research topics, identify the strategic direction for future works, and shape the culture of this growing community [1]. For research studies, using proper keywords will improve publication visibility in research networks. Publication visibility is defined as the share of traffic a study receives based on its ranking from Search Engine Result Pages (SERPs), which can be gained by indexing the relevant keywords to help search engines find the related studies [2]. It is essential to acknowledge that analyzing keyword selections in the context provided by the publication abstract enables us to determine the keywords' relevance to the research topics. However, evaluating the selection of publication keywords is an essential but lesser-addressed issue [3].

In this work, we discuss how examining bibliometric data can be a practical approach to scrutinizing publication keywords to bring conceptual and functional insights into CSER's domain. The research questions guiding this work are:

1. *What are the most prevalent research foci represented in CSER through publication keywords?*
2. *How well do publication keywords align with the CSER publications' stated strategic directions?*
3. *How well do selected keywords in CSER publications adequately represent the research topics presented via publication abstracts?*

We gathered the titles, abstracts, and keywords collectively from studies published in two major computer science (CS) education conferences and journals from 2015 to 2019 to answer these questions. We also extracted the strategic directions from their mission statements. By applying Term Frequency-Inverse Document Frequency (tf-idf) to the textual data, we were able to not only identify the most prevalent research foci represented by the most commonly used keywords of CSER, but also to examine whether those chosen keywords are in alignment with the corresponding strategic directions and abstract content.

2 Literature Review

This literature review discusses the background of the research publication keywords and visibility in subsection 2.1. Then, we cover the bibliometrics analysis approach and its application in education research in subsection 2.2.

2.1 Publication Keywords and Visibility

Publication keywords are considered as essential elements of representing knowledge concepts and have been widely utilized to reveal research domains' knowledge structure [4]. It ensures that the research paper is properly indexed by databases and search engines improving the research visibility [5]. With the growing number of publications available through the internet, many keyword extraction approaches have been studied and explored, including without limitation, automatic keyword extraction from abstract [6], [7], machine learning-based approaches for keyword extraction [8]. Most of these studies are in the research discipline of the information sciences or computer sciences to study effective information retrieval techniques focusing on keywords and their application. In bibliometric research, many studies have explored keywords to identify research topics and interpret the results at both macro-level to demonstrate the structural characteristics of domain knowledge and micro-level to reveal the details of a domain's research topics and their relations [3]. Our work sought to extend the studies of keywords to CSER discipline by leveraging the opportunities demonstrated through a bibliometric approach.

2.2 Bibliometric Analysis and Its Application in Education Research

Bibliometric analysis is a statistical evaluation of published scientific articles, books, or chapters of a book. It is an effective way to measure a publication's influence in the scientific community [9, 10]. Bibliometrics can be used to provide evidence of research impact, find new and emerging areas of research, identify potential collaborators, and recognize relevant or high-impact journals [11]. While it has been used in various research domains, this paper only discusses its application in education research.

In education research, bibliometric analysis has been applied to measure research performance and characteristics in various domains, including mathematics education [12, 13], higher education in the UK [14]], China [15] and Switzerland [16], doctoral education [17], among others. There is science education [18, 19] or STEM education [20] research that has deployed bibliometric approaches to identifying trends, and limited studies can be found in computer science education. Marti-Parreno et al. [21] have explored gamification trends in education. Xian & Madhavan [22] have examined scholarly collaboration in engineering education. Cheng et al. [23] have discovered research themes for e-learning in the workplace. Shen and Ho [24]] have investigated technology-enhanced learning research trends by identifying the 40 most influential articles and their authors in the field. These studies are in computer science education-related domains, but none of them

directly addressed computer science education research with a bibliometric approach. Papamitsiou et al.[25] have applied co-word analysis with social network analysis to keywords from two conferences' publications in computing education. This bibliometric study characterized the CSER research landscape by showing the dominant research fields: learning approaches, aspects of programming, computational thinking, feedback, and assessment. Integrity and diversity are two additional areas that attracted researchers' attention. Merlo et al. [26] have used Lotka's law, an empirical law used in bibliometric studies, in analyzing research trends in computer science education publications in 2011. However, that study was focused on the frequency of topics by authors in CSER to conclude trends and future paths. As an effective tool to discover knowledge, the bibliometric analysis approach has yet to be actively applied to the growing field of CSER. Our work took further steps by combining the NLP techniques with the bibliometric approach to better facilitate the process to find insights. In addition, we explored adopting a marketing framework of Segmentation, Targeting, and Positioning to frame our study.

3 Theoretical Framework

By evaluating the keyword selections in the CSER research publications, we aim to generate insights to improve publication visibility. Similar to business organizations pursuing to achieve their key performance indicators (KPIs), researchers' impacts are commonly evaluated through bibliometrics. Bibliometrics is one KPI for research impact. We argue that scholars and researchers are like business organizations, their research publications are like products, and their audiences are like customers in the market. Metaphorically, the goal of improving publication visibility is a similar process to finding the product-market fit that can optimize the market potential. With these concepts, we applied a marketing framework for Segmentation, Targeting, and Positioning (STP) to guide this study.

STP is a framework that segments the market, targets selected segments with marketing resources tailored to the market's preferences, and adjusts positioning accordingly [27]. It is an empirical approach that focuses on breaking the market into smaller segments, so as to allow developing specific strategies to reach and engage the audience [28]. We applied STP to guide this work that each of the research questions corresponds to one paradigm of this framework. Market segmentation determines the critical characteristics of the market [27]. In our study, we reviewed the authors' keywords as a base to obtain the segments of CSER. Market targeting is to evaluate the attractiveness of the characteristics of each segment [27]. The strategic directions of CSER publication venues are considered the targeted segments in CSER. The keywords selected for publication represent the existing segments of CSER. By examining the strategic directions and the selected keywords, we might be able to determine the target segments with high attractiveness if we discover an alignment. Market positioning is the development of the market mix to reach and engage the audience [27]. We examined the potential for publication to reach and engage the audience through evaluating the alignment between authors' selected keywords and abstract generated keywords. Selecting the most salient keywords can significantly increase the chances of a document being retrieved by the publication's pertinent readers and promote an article's visibility [29]. An abstract is a guide to the essential parts of the manuscript, and many researchers will only read the abstract of a publication [30]. We argue that when the authors' selected keywords align with the abstract generated keywords, it will reinforce the selected keywords' topic relevance. Consequently, it will decrease the attrition for the pertinent audience to locate such publication.

4 Methods

In this part, we present the overall summary of the research design in subsection 4.1. Subsection 4.2 addresses how we create the sample of the study through web scraping and how to collect and process the dataset. Subsection 4.3 discusses how we analyze textual data by using NLP.

4.1 Research Design

This study aims to conduct keyword analysis in computer science education research using a bibliometric approach. An overview of the mixed-method research design, including obtaining data from the publication database, data processing with the textual data, and data analysis techniques to obtain valuable insights, is presented in Figure 1 below. First, we applied web scraping techniques to collect the keywords from the publication database. Web scraping is a practical approach to transform unstructured data on the web into structured data that can be stored and analyzed in a central local database or spreadsheet [31]. With this approach, we collected the articles' publication information, including titles, publication date, keywords, and abstracts, to answer the research questions for this study. Second, we applied data processing to the structured data scraped from the publication database. Data processing transforms raw text into usable forms to perform analysis and generate insights [32]. Third, we performed data analysis via NLP techniques. NLP can be useful for rapidly extracting codes from text and can support qualitative data analysis [33]. This study's analyzed results will then be visualized for reporting and validated to ensure the reliability of the automated analysis process powered by NLP. The above tasks for this study were performed by Python 3.7.6.

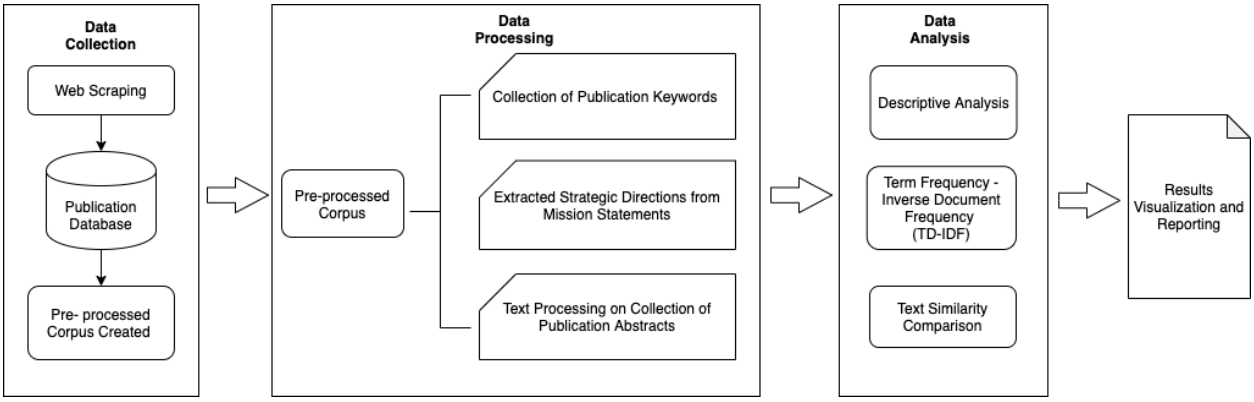

Figure 1: Overview of the research design

4.2 Data Collection and Processing

To select a representative sample for our study, we collected articles published from two major CS education conferences and journals from 2015 to 2019. After pre-processing data to remove duplicates and unnecessary fields, we gathered articles' titles, paper URLs, keywords, and abstracts for further analysis. We used Python BeautifulSoup [34] and Requests [35] libraries to scrape the publication database to generate the dataset. We utilized the metadata tags on each publication URL to identify the correct tags for data collection. By examining the scraped dataset, we found missing values of specific records due to incorrect or missing keywords information provided in the metadata "keywords" tags. By further investigating the missing values, we either filled in the missing values to the extent we could identify keywords, or removed the records from the dataset in case we could not locate such information [36]. We then performed data processing and ended

with n = 1,775 records for the next step in our analysis. We applied the following Python libraries and packages to process the data:

- **Pandas**: an open-source data analysis and manipulation tool for Python [37]. We used version 1.1.0 in the analysis.
- **Numpy**: an open-source project to enable numerical computing in Python [38]. We applied version 1.18.1.
- **lxml**: a Python library for processing XML and HTML [39], and we used version 4.3.0 in data collection and processing.
- **Natural Language Toolkit (NLTK)**: a suite of open-source libraries for symbolic and statistical natural language processing for English written in Python [40]. We used version 3.3.

Data collection started in early October 2020 and completed in mid-November 2020. Since the conference proceedings of 2020 have not been published yet at the time of the data collection, we ended up collecting the conference proceedings from 2015 to 2019, a span of five years. For all the records in the dataset, we parsed the "keywords" column by applying the Python NLTK library [40] to prepare the text data for frequency computation. To better capture detailed CSER segments, six domain-specific keywords were removed to segment the domain out of the main topics, so as to reduce the information noise and obtain the content-focused keywords for further analysis. After processing the removal , the final keywords list contained 4,646 unique keyword values. All keywords were formatted in lowercase for accurate analysis [41]. To extract the strategic directions from mission statements, we coded the text of mission statements with related themes, which were used to compare with the publication keywords list. To prepare the abstract for text analysis, we applied the NLTK library to convert abstracts into lowercase and to remove punctuation, special characters, and stopwords for any computational-based NLP tasks [41].

4.3 Data Analysis

There were three data analysis parts that used NLP techniques, with each corresponding to the research questions defined for this study. The first part was to analyze keywords collected from the publication metadata to identify the represented research foci. The second part was to compare those collected publication keywords with the strategic themes generated from the mission statements. The last part of the analysis explored the alignment between collected publication keywords and the abstract generated keywords.

The word frequencies were calculated by using the processed unique keywords list based on the Python Collections library [42]. The commonly used keywords were those with the highest word frequencies representing CSER's research foci. The top 1,000 most frequently used words were generated along with their frequency counts. The results were visualized using Python libraries WordCloud [43] and Matplotlib [44] for reporting. The keywords list with the highest frequencies was compared with the extracted strategic directions to determine whether there is a mismatch between CSER's research foci and the strategic directions with the academic publications. The authors of the papers often selected the publication keywords based on their work knowledge, mainly using generic terms as they reflect a rough overview of a scientific discipline or represent popular themes [3]. To investigate whether the keywords chosen by authors align with the topics

represented through the paper abstracts, we explored the NLP technique named Term Frequency-Inverse Document Frequency (tf-idf).

Tf-idf [45] is a method that identifies important terms by the product of two statistics, term frequency and inverse document frequency. It is intended to reflect how important a word is to a document in a collection or corpus [46]. By applying the tf-idf to the abstract, a topic relevance word list was generated based on each abstract's tf-idf ranking. We compared this list with the keywords selected by the authors to determine whether the chosen keywords adequately represent the research topics presented through abstracts for each research publication. For comparisons between the authors' selected keywords list and the topic relevance keywords list generated from the abstract, Ratcliff-Obershelp Pattern Recognition (also known as Gestalt Pattern Matching) was implemented. It is a sequential-based algorithm that can identify the longest common sequence in the strings [47]. The longer the common sequence is identified, the higher similarity of the two strings can be found [48]. We utilized Python Textdistance, a library for comparing the distance between two or more sequences by many algorithms [49]. For both lists, keywords were stored in text format as strings in the data frame. A string is a type of data in computer programming used to store a sequence of elements, typically characters [50]. Thus, we can apply this algorithm to identify similarities of keywords strings by computing the number of matching characters in one string divided by the total number of characters in two strings [47].

We explored a manual approach to inspect the validity of the results from the automated analysis process. Applying a similar approach to a prior study [51], a randomly selected subset with 50 records was created for validation. We examined the keywords generated from the abstract that presented the abstract context through tf-idf. Additionally, we checked the accuracy of string similarity calculation by manually inspecting the author-selected keywords and the abstract-generated keywords in the subset.

5 Results

A total of 4,646 unique keywords selected by the authors were collected from n = 1,775 records. The average keywords selected by the authors were 4.44 words per publication. As mentioned in section 4, the domain-specific keywords were removed from the unique keywords' list to stay with the topic relevance approach. The word frequency counts were computed, with the top 25 keywords selected by authors presented in Figure 2 below. As shown in the figure, five most frequent keywords were "project-based learning," "assessment," "active learning," "STEM," and "motivation." The keywords with the highest frequencies also showed that most of the work from 2015 to 2019 was in teaching and learning areas, represented by keywords such as "project-based learning," "active learning," "motivation," "programming," and "software engineering." Broadening participation is another area that caught CSER researchers' attention with keywords mostly selected, including but not limited to "diversity," "K12," "first-year engineering," and "engineering identity."

A word cloud visualization generated based on clustering the most frequently selected keywords is displayed in Figure 3. It revealed similar themes for teaching & learning and broadening participation. "Learning," "programming," and "assessment" are specific keywords which represent the word cloud's teaching & learning theme. As to broadening participation, keywords under this theme include "diversity," "first year," and "identity." This result supported the identified trend

for prevalent research foci that teaching & learning and broadening participation were the most common topics researched in CSER.

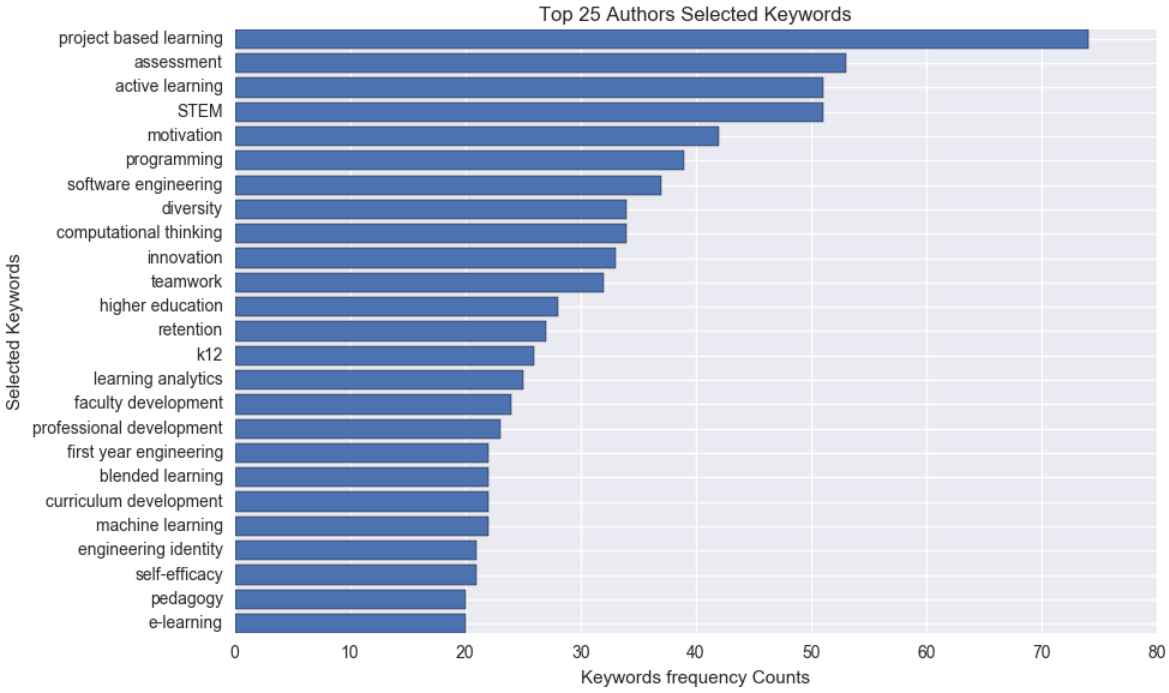

Figure 2: Top 25 Authors' Selected Keywords

Figure 3: Word Cloud of Author' Selected Keywords

Furthermore, by examining CSER academic strategic themes from which data was collected, the two themes - teaching & learning and broadening participation - were also represented in the corresponding strategic directions. The strategic themes were determined by coding the mission and value statements of the two academic venues manually. We need to further establish the validity of theme coding through member checking or other appropriate techniques [52]. The current analysis showed that the research foci align with the strategic directions in the CSER domain.

The topics of abstracts were represented by the keywords generated through Term Frequency-Inverse Document Frequency (tf-idf) method. The abstract texts were processed and stored in dataframe format in Python for math computation. Through a quick exploratory analysis of the authors' selected keywords, the average keywords chosen by authors for indexing were 4.44. With this information in mind, we applied tf-idf to the abstract to obtain five feature words as the extracted abstract keywords. We computed the sequential distance using Ratcliff-Obershelp Pattern Recognition by identifying the longest common sequence of two strings to compare the authors' keywords with the abstract keywords. As noted in section 4.3, the computed values indicate the likelihood of the two strings' similarity, which falls between 0 to 1 [48]. Our results showed that the similarity values between the two keyword lists ranged from 0.06 to 0.78, with a mean value of 0.32. Prior work using NLP techniques commonly set a similarity threshold that is greater than 0.7 [53–55]. In our study, only one record was found with a similarity greater than 0.7. We inspected the percentage of documents that met the threshold range from 0.4 to 0.6 with 0.1 intervals. As noted previously, an average of 4.44 keywords was used to index one paper. A 0.4 similarity threshold means that the two strings of keywords - strings for the authors' keywords and strings for abstract extracted keywords - have more than one keyword in common. A greater than 0.5 similarity threshold suggests that at least half of the words in both keyword lists are similar. Table 1 shows our analysis that only 4.96% of the total records had keyword similarity higher than 0.5, indicating a misalignment between the selected keywords and the abstract generated keywords.

Table 1: Authors Keywords and Extracted Keywords Similarities

Keywords Similarity Threshold	Percentage of documents meet the threshold
Greater than 0.4	22.54%
Greater than 0.5	4.96%
Greater than 0.6	0.68%

6 Discussion

The results of our analysis suggested that the most prevalent research foci represented by the most commonly used keywords of CSER were on teaching & learning and broadening participation, which aligned with the corresponding strategic directions between 2015 and 2019. This finding was consistent with the prior work performed by Papamitsiou et al. [25] using data from two different publication venues in computing education. With increased attention to broadening participation, our analysis suggested that the CSER community has been putting efforts into formulating a diversified culture through teaching and learning for inclusivity. However, the research foci were drawn from the top frequency of keywords selected by the authors. This method was only based on the word counts in the document context which might neglect the keywords' semantic meanings and associated contexts. A keyword semantic-based clustering approach to identify the research foci themes will be needed for future work. From the top 25 keywords selected by the authors during the span of 5 years, there are broad terms, like "STEM," that we could not simply state a theme associated with it. Further analysis, such as association analysis or network analysis, must be undertaken before confirming such an association. For example, if "STEM" is connected more to "programming" or "software engineering," it would more likely be related to teaching & learning. On the other hand, if "STEM" is more associated with "diversity" or " first-year engineering," it would likely be related to broadening participation. Also, various levels of connections in a

network analysis may indicate different priorities of the associated contexts. For future research, it will be essential to further our knowledge by implementing association analysis or network analysis along with the bibliometric data to study the contexts and priorities of the keywords.

Moreover, authors select keywords based on their knowledge and experience with their work. Even though guidelines on determining the keywords for the research publication are provided to authors, nuances of choosing the keywords still exist. For example, for the keyword "innovation," variations of keywords selected by authors include "innovation and creativity," "innovative pedagogy," "innovation in education," "innovative curriculum," and "educational innovation." As indicated in prior studies [3, 25], authors tend to choose more generic and higher-level terms in selecting keywords, such as "innovation" or "innovation and creativity." They may also opt to use concrete terms with more details, such as "innovative pedagogy" or " innovative curriculum." Choosing specific keywords with more semantic meanings can be a double-edged sword. It might have a negative impact on publication's visibility, as suggested by the STP framework, unless the exact and specific terms are used in the scholarly literature search by the pertinent audience. To investigate this issue in the future, an in-depth nuanced analysis will be necessary and helpful. One approach to further this work is to generate semantic-based clustering keyword lists. We also believe that manually coding keywords list into groups with related themes will benefit the nuanced analysis to evaluate further and validate the research foci.

As to the keywords extraction from the abstracts, similar semantic-based issues remain. The current results reported a misalignment between the authors' selected keywords and the topic relevance abstract generated keywords through the tf-idf method. Our keyword extraction is based on the assumption that the higher the word's frequency in a document, the greater likelihood of it being selected as the keyword (feature) for its text. A similar drawback of this approach is the semantic context. For future work, other natural language processing techniques for semantic-based keyword extraction methods should be explored.

Meanwhile, our work has demonstrated how combining the bibliometrics approach and NLP can further knowledge in CSER. Using a separate dataset in our study reinforced the insights discovered from prior work which took a similar approach to keywords [25]. Besides the results generated to address the defined research questions, the data collection processing revealed certain findings that are worthy of notice. Authors usually define keywords in their manuscripts and submit them via the designated publication portal. Manuscripts are then indexed to publish and stored in metadata, available via the publishers' websites. We searched for each paper's metadata via the publication venues' sites when we collected paper data. Then the tags in the metadata were used to collect the relevant data fields about each paper. In our process, some papers could not identify their keywords due to errors found in the corresponding keywords tags of metadata used for indexing. We would recommend authors double-check their published manuscript through the metadata on the publisher's website for better visibility of their research publications.

Lastly, we would recommend practical strategies to improve publication visibility based on the findings of our study. As Google has been one of the major search engines that drive scholarly search traffic [56], authors can leverage keywords with some searching engine optimization (SEO) techniques to optimize their publication search results. The followings are some commonly suggested SEO strategies listed under authors' resources by publishers [30, 56] or libraries [11, 57]. The first SEO strategy for authors to consider is to include at least one of the selected keywords

in the title to be better captured by searching algorithm. As for abstract, since the first three lines will be displayed for each of the Google Scholar search results, authors should include essential components of the study and keywords in the first two sentences to make it SEO-friendly. The importance of keyword selection cannot be overstated in the context of publication visibility and SEO. One of the techniques to select the publication keywords, as our results implied, is to think about the tf-idf of the abstract to generate keywords. When choosing one specific keyword from two or more similar ones, authors may want to evaluate those candidate keywords through the Google Trends keyword tool. Of note, some publication venues require authors to follow a specific keywords classification system. With the rapid growth of CSER, it is not easy to fit certain emerging keywords into a system developed years ago. We want to encourage the CSER community to re-evaluate the current keyword lists used for indexing publication on a regular basis and adapt to the emerging trends to better support the growth of this field.

7 Conclusion

In summary, our analysis provides answers to the three research questions defined for this study. We have demonstrated how applying a marketing theoretical framework can further understand publication visibility and research impact in CSER. We have also exemplified how functional insights can be discovered via the combination of the bibliometrics analysis approach and natural language processing techniques. We hope our suggestions motivate scholars and researchers to carefully evaluate and select keywords for indexing publications to improve the research topic relevancy and publication visibility for broader impact.

References

[1] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS quarterly*, vol. xiii-xxiii, 2002.

[2] A. Ortiz-Cordova and B. J. Jansen, "Classifying web search queries in order to identify high revenue generating customers," *Journal of the American Society for Information Sciences and Technology*, vol. 63, no. 7, pp. 1426–1441, 2012.

[3] G. Chen and L. Xiao, "Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods," *Journal of Informetrics*, vol. 10, no. 1, pp. 212–223, 2016.

[4] H. N. Su and P. C. Lee, "Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight," *Scientometrics*, vol. 85, no. 1, pp. 65–79, 2010.

[5] V. Rodrigues, "How to write an effective title and abstract and choose appropriate keywords," *Editage Insights (04-11-2013)*, 2013.

[6] Y. HaCohen-Kerner, "Automatic extraction of keywords from abstracts," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2003, pp. 843–849.

[7] Y. B. Wu, Q. Li, R. S. Bot, and X. Chen, "Domain-specific keyphrase extraction," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 283–284.

[8] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," in *international conference on web-age information management*. Springer, 2006, pp. 85–96.

[9] A. Schubert, W. Glänzel, and T. Braun, "Scientometric datafiles. a comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981–1985," *Scientometrics*, vol. 16, no. 1-6, pp. 3–478, 1989.

[10] F. de Moya-Anegón, Z. Chinchilla-Rodríguez, B. Vargas-Quesada, E. Corera-Álvarez, F. Muñoz-Fernández, A. González-Molina, and V. Herrero-Solana, "Coverage analysis of scopus: A journal metric approach," *Scientometrics*, vol. 73, no. 1, pp. 53–78, 2007.

[11] "Measuring research impact." [Online]. Available: https://library.leeds.ac.uk/info/1406/researcher_support/17/measuring_research_impact

[12] Q. Chen and P. Tian, "Bibliometrics analysis of journal of mathematics education in 2008 - 2010," *Journal of Mathematics Education*, vol. 4, 2011.

[13] P. Drijvers, S. Grauwin, and L. Trouche, "When bibliometrics met mathematics education research: the case of instrumental orchestration," *ZDM*, pp. 1–15, 2020.

[14] J. Adams, "The use of bibliometrics to measure research quality in uk higher education institutions," *Archivum immunologiae et therapiae experimentalis*, vol. 57, pp. 19–32, 02 2009.

[15] W. Y. W. Haiyan, "Analysis of academic characteristic of the highly cited papers in chinese higher education field-based on the bibliometrics of literatures from china higher education research (2000-2011)[j]," *China Higher Education Research*, vol. 1, 2012.

[16] A. Diem and S. C. Wolter, "The use of bibliometrics to measure research performance in education sciences," *Research in higher education*, vol. 54, no. 1, pp. 86–114, 2013.

[17] P. Macauley*, T. Evans, M. Pearson, and K. Tregenza, "Using digital data and bibliometric analysis for researching doctoral education," *Higher Education Research & Development*, vol. 24, no. 2, pp. 189–199, 2005.

[18] F. Arici, P. Yildirim, Şeyma Caliklar, and R. M. Yilmaz, "Research trends in the use of augmented reality in science education: Content and bibliometric mapping analysis," *Computers & Education*, vol. 142, p. 103647, 2019.

[19] V. Larivière, C. Ni, Y. Gingras, B. Cronin, and C. R. Sugimoto, "Bibliometrics: Global gender disparities in science," *Nature News*, vol. 504, no. 7479, p. 211, 2013.

[20] S. G. Assefa and A. Rorissa, "A bibliometric mapping of the structure of stem education using co-word analysis," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 12, pp. 2513–2536, 2013.

[21] J. Marti-Parreño, E. Méndez-Ibáñez, and A. Alonso-Arroyo, "The use of gamification in education: A bibliometric and text mining analysis," *Journal of Computer Assisted Learning*, vol. 32, 08 2016.

[22] H. Xian and K. Madhavan, "Anatomy of scholarly collaboration in engineering education: a

big-data bibliometric analysis," *Journal of Engineering Education*, vol. 103, no. 3, pp. 486–514, 2014.

[23] B. Cheng, M. Wang, A. I. Mørch, N. S. Chen, and J. M. Spector, "Research on e-learning in the workplace 2000–2012: a bibliometric analysis of the literature," *Educational research review*, vol. 11, pp. 56–72, 2014.

[24] C. W. Shen and J. T. Ho, "Technology-enhanced learning in higher education: A bibliometric analysis with latent semantic approach," *Computers in Human Behavior*, vol. 104, no. 106177, 2020.

[25] Z. Papamitsiou, S. Michail Giannakos, and A. LuxtonReilly, *Computing Education Research Landscape through an Analysis of Keywords*. August 10–12 Virtual Event, New Zealand. ACM, New York, NY, USA, 11 pages: In Proceedings of the 2020 International Computing Education Research Conference (ICER '20), 2020. [Online]. Available: https://doi.org/10.1145/3372782.3406276

[26] C. R. Merlo, J. M. Merlo, L. Hoeffner, and R. Moscatelli, "An analysis of computer science education publication using lotka's law," *Journal of Computing Sciences in Colleges*, vol. 26, no. 3, pp. 85–92, 2011.

[27] J. T. Bowen, "Market segmentation in hospitality research: no longer a sequential process," *International Journal of Contemporary Hospitality Management*, vol. 10, no. 7, pp. 289–296, 1998. [Online]. Available: https://doi.org/10.1108/09596119810240924

[28] D. Chaffey and D. Bosomworth, *Digital Marketing: Strategy*, 2012.

[29] D. C. F. Garcia, C. C. Gattaz, and N. C. Gattaz, "The relevance of title, abstract and keywords for scientific paper writing," *Revista de Administração Contemporânea*, vol. 23, no. 3, pp. 1–9, June 2019. [Online]. Available: https://doi.org/10.1590/1982-7849rac2019190178

[30] "Title, abstract and keywords." [Online]. Available: https://www.springer.com/gp/authors-editors/authorandreviewertutorials/writing-a-journal-manuscript/title-abstract-and-keywords/10285522

[31] A. V. Saurkar, K. G. Pathare, and S. A. Gode, "An overview on web scraping techniques and tools," 2018, international Journal on Future Revolution in Computer Science & Communication Engineering, 4(4), 363-367.

[32] Sciforce, "Text preprocessing for nlp and machine learning tasks," May 2020. [Online]. Available: https://medium.com/sciforce/text-preprocessing-for-nlp-and-machine-learning-tasks-3e077aa4946e

[33] A. E. E. . H. R. Crowston, K., "Using natural language processing technology for qualitative data analysis," *International Journal of Social Research Methodology*, vol. 15, no. 6, p. pp. 523–543, 2012.

[34] L. Richardson. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/

[35] "Requsts: Http for humans™." [Online]. Available: https://requests.readthedocs.io/en/master/

[36] Dansbecker, "Handling missing values." vol. 26, February 2019. [Online]. Available: https://www.kaggle.com/dansbecker/handling-missing-values

[37] W. McKinney, "Data structures for statistical computing in python," *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51–56, June 2010.

[38] C. Harris, K. Millman, and et al., "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[39] S. Behnel, M. Faassen, and I. Bicking, *lxml: XML and HTML with Python*, 2005.

[40] S. Bird, "Nltk: the natural language toolkit," *In Proceedings of the COLING/ACL Interactive Presentation Sessions*, pp. 69–72, July 2006.

[41] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media,Inc., 2009.

[42] F. Lundh, *Python Standard Library.* O'Reilly Media,Inc., 2001.

[43] "Wordcloud for python documentation." [Online]. Available: https://amueller.github.io/word_cloud/

[44] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[45] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[46] A. Rajaraman and J. D. Ullman, "Data mining (pdf)," *Mining of Massive Datasets*, no. ISBN 978-1-139-05845-2., p. pp. 1–17., 2011.

[47] P. E. Black, "Ratcliff/obershelp pattern recognition." [Online]. Available: https://www.nist.gov/dads/HTML/ratcliffObershelp.html

[48] M. Mayank, "String similarity - the basic know your algorithms guide!" Jan 2020. [Online]. Available: https://itnext.io/string-similarity-the-basic-know-your-algorithms-guide-3de3d7346227

[49] Orsinium, "textdistance." [Online]. Available: https://pypi.org/project/textdistance/description

[50] "String (computer science)," Jan 2021. [Online]. Available: https://en.wikipedia.org/wiki/String_(computer_science)

[51] S. Lunn, J. Zhu, and M. Ross, "Utilizing web scraping and natural language processing to better inform pedagogical practice," *2020 IEEE Frontiers in Education Conference (FIE),Uppsala,*, pp. 1–9, 2020.

[52] D. Yanow and P. Schwartz-Shea, *Interpretation and method: Empirical research methods and the interpretive turn*. Routledge, 2015.

[53] B. Ong, R. Wen, and A. N. Zhang, "Data blending in manufacturing and supply chains," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 3773–3778.

[54] M. S. Maučec, Z. Kačič, and B. Horvat, "Modelling highly inflected languages," *Information Sciences*, vol. 166, no. 1-4, pp. 249–269, 2004.

[55] Intellica.AI, "Comparison of different word embeddings on text similarity-a use case in nlp," Oct 2019. [Online]. Available: https://intellica-ai.medium.com/comparison-of-different-word-embeddings-on-text-similarity-a-use-case-in-nlp-e83e08469c1c

[56] "Search engine optimization (seo) for your article." [Online]. Available: https://authorservices.wiley.com/author-resources/Journal-Authors/Prepare/writing-for-seo.html

[57] "Research visibility: Seo for authors: A how-to guide." [Online]. Available: https://guides.library.ucla.edu/seo/author