POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Using text mining for research trends in empirical software engineering

Deneysel yazılım mühendisliğindeki araştırma eğilimleri için metin madenciliği

***Yazar(lar) (Author(s))**: Gül TOKDEMİR*

ORCID: 0000-0003-2441-3056

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Tokdemir G., "Using text mining for research trends in empirical software engineering ", ***Politeknik Dergisi***, 24(3): 1227-1235, (2021).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.831391**

Using Text Mining for Research Trends in Empirical Software Engineering

Highlights

- *The analysis has revealed that there is an increasing interest to Empirical Software Engineering research, especially in the last two decades*
- *For the last 20-year period three topics, namely, Organization, Software Development Process and Human, got the most attention from researchers in Empirical Software Engineering research.*
- *The fastest growing research topics are found to be Performance, Coding, Testing and Software Practice topics based on the two decades comparisons.*

Graphical Abstract

This paper aims to examine the research trends in Empirical Software Engineering domain by applying topic modelling.

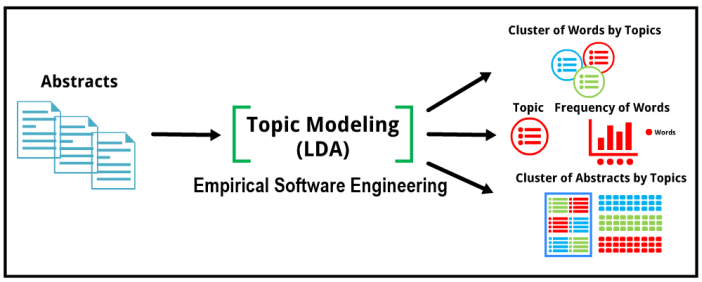


Figure. Topic Modeling for research trends in Empirical Software Engineering

Aim

This paper intends to investigate and analzye the research trends in Empirical Software Engineering domain within the last two decades using text mining.

Design & Methodology

The abstracts of 10658 articles published in the literature were analyzed using topic modeling, and the research topics in this field were investigated and analyzed comparatively.

Originality

This study is the first in applying the topic modeling technique in Emprical Software Engineering literature.

Findings

The analysis for the last 20-year period have shown that; Organization, Software Development Process and Human topics got the most attention from researchers and the fastest growing research topics are Performance, Coding, Testing and Software Practice.

Conclusion

It can be expected that the Empirical Software Engineering research will gain more interest from the research community as software industry needs for more evidence on the practical applications of the theory.

Declaration of Ethical Standards

The author of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Deneysel Yazılım Mühendisliğindeki Araştırma Eğilimleri için Metin Madenciliği

Araştırma Makalesi / Research Article

Gül TOKDEMİR

Bilgisayar Mühendisliği Bölümü, Çankaya Üniversitesi, Ankara, Türkiye

(Geliş/Received : 25.11.2020 ; Kabul/Accepted : 26.02.2021 ; Erken Görünüm/Early View : 29.03.2021)

ÖZ

Bu çalışma Deneysel Yazılım Mühendisliği alanında son yirmi yılda ki araştırma eğilimlerini metin madenciliği tekniklerini kullanarak incelemeyi amaçlamaktadır. Makale özetleri göz önünde bulundurularak, Deneysel Yazılım Mühendisliği ile ilgili literatürde yayınlanmış 10658 makale incelenmiştir. İstatistiksel bir modelleme tekniği olan (Latent Dirichlet Allocation) kullanılarak, bu alandaki temel araştırma konuları bulunarak karşılaştırılmalı olarak incelenmiştir. Bu makalede son yirmi yıl içinde yayınlanmış çalışmalardaki odak değişiklikleri değerlendirilmekte ve araştırma içeriğindeki son eğilimler ortaya çıkarılmaktadır. Karşılaştırmalı değerlendirme yoluyla, deneysel yazılım mühendisliği alanındaki araştırma eğilim değişikliği vurgulanarak, hem akademisyenler hem de uygulayıcılar için faydalı olabilecek ve bu alanın ilerlemesini sağlayacak araştırma gündemi önerilmektedir.

Anahtar Kelimeler: Deneysel yazılım mühendisliği, konu modelleme, araştırma eğilimleri, latent dirichlet allocation.

Using Text Mining For Research Trends in Empirical Software Engineering

ABSTRACT

This paper intends to examine the research trends in Empirical Software Engineering domain within the last two decades using text mining. It studies published articles in the relevant literature with an emphasis on abstracts of 10658 articles published in the literature on Experimental Software Engineering domain. Using a probabilistic topic modelling technique (Latent Dirichlet Allocation), it brings forward the main topics of research within this domain. By further analysis, the paper evaluates the changes of focus in published works in the last two decades and depicts the recent trends in research content wise. Through a timely comparison, it portrays the alteration of interest within empirical software engineering research and proposes a future research agenda to develop an advanced field, beneficial both for academics and practitioners.

Keywords: Empirical software engineering, topic modelling, research trends, latent dirichlet allocation.

1. INTRODUCTION

There has been an increased interest in the software engineering, especially within the last twenty years, as a result of accelerated technological developments and progressive amount of innovative works worldwide. Accordingly, there has been a rise in the number of empirical investigations in the software engineering domain and these studies have diversified rapidly, through demands and needs of industry and academia, in parallel. This situation also revealed itself in the relevant literature as a surge of published work [1], i.e. articles in journals, conference papers/proceedings, extending the boundaries of empirical work in the software engineering. Experiments and case studies have been mostly preferred in empirical studies, besides other methods like correlational studies, meta-analysis and questionnaires [2]. Especially in academia, use of experiments has been a common research method, with the use of students, whereas within the industry, there has been a disinclination towards running experiments, leaving the situation blurred and problematic for software engineering community [3].

*Corresponding Author
e-mail : gtokdemir@cankaya.edu.tr

It is acknowledged that empirical studies in software engineering have a crucial part in the advancement of software engineering discipline [4], i.e. developing sound and practical theories [5], with an influence on the practice of software engineering [6] consequently. However, today, the literature is still fragmented and incoherent due to lack of longitudinal studies, absence and/or underrepresentation of important topics [2], a low number of in-depth research and fewer important works in academia, restricting a detailed and widened understanding of concepts related to software engineering. In this sense, creating guidelines in empirical studies to improve the poor standards of software engineering research and to construct a new perspective in empirical practices is crucial [7]. In a similar way, more empirical research, with greater quality and importance, and additional focus on blended research and theory construction are required in future works [8].

In this sense, a systemic review is also considered as a part of theory building in empirical studies. MacDonell et al. [9] consider it as a useful and a common research instrument tool. Study of Mathew et al. [10] emphasizes that using text mining in systemic review, large scale

trends in empirical research in contemporary software engineering can be detected. In terms of bibliographic search, also a step in our study, the importance of specifying the bibliographic search conducted for a systematic review is considered important [11], whereas the need for "a comprehensive bibliometric assessment" for a better understanding of the vast and dynamic literature is argued [1].

In the following sections, we present our research methodology; topic modeling based on Latent Dirichlet Allocation (LDA), a probabilistic text mining technique. We provide the steps of this process in detail and present the outcomes in tables and graphs. We discuss and evaluate the findings; research topics and their changes, summarize the trends and alterations of interest in empirical studies and propose a future research agenda for academics and practitioners.

2. BACKGROUND

Empirical software engineering (EmpSWE) focuses on qualitative and quantitative scientific measurement of both software engineering process and product [5] which endorses empirical evidence as the major knowledge source [12]. It promotes the use of empirical methods to understand software development process and software product in a better way [13]. It aims to create a systematic understanding of effectiveness of software engineering technologies on different actors while performing various activities on different software systems [8]. Sjoberg and Dyba [8] envisioned that in years 2020-2025, software engineering research that proposes new or modified technology would be supported with empirical evidence for publication more and big software companies would keep skilled personnel to conduct such empirical studies.

Topic Modeling has a wide range of application in various domains such as hotel online reviews [14], newspaper archives [15], linguistics, [16][17], finance [18]. Many algorithms such as NMF (Non-negative Matrix Factorization), PCA (Principal Component Analysis), LSA (Latent Semantic Analysis), RP (Random Projections), LDA have been applied to reveal topics for different problem domains. The performances of these algorithms and their advantage and disadvantages are explored [18]-[20]. LSA (Latent Semantic Analysis), one of these algorithms, offers the advantages of catching the synonyms of words and by solving the problem of data sparseness. However, this algorithm cannot provide a statistical basis for determining the number of topics. Although NMF performs fast processing for large-scale data, it could give semantically incorrect results. PCA performs very well in low dimensional data, but it is difficult to evaluate the covariance matrix correctly, accordingly an effective analysis for high dimensional data sets cannot be achieved. RP algorithm can be used strongly in unbalanced data sets, yet it is effective mainly in analysis of small data sets. LDA can be applied to long documents, providing semantically interpretable results. However, the algorithm cannot model the relationships between the created topics.

LDA is the most popular algorithm which is studied extensively in many domains [21] and reported to work effectively in generating the context for the collection of documents [22]. [23] overviews topic modeling approaches and suggest a guideline for choosing most suitable method for the specific analysis. By following the guidelines of [23], for the scientific paper abstract's analysis where the topic relationships are not in the scope of the analysis, LDA method was chosen for this study. LDA-based topic modeling has been applied in software engineering field for different purposes as it is applied in many other fields [24]-[28]. The LDA algorithm infers topics from repetitive patterns of word existences [19]. It uses mixed membership approach [29] where a document is considered to contain several topics.

In the review study of Höfer and Tichy [2], empirical research in the software engineering domain between 1996-2006 is explored and they conclude that the variety of topics studied were quite narrow which should be widened. In this study, using topic modelling, we explore the recent trends in EmpSWE research, identifying the main topics of study and the changes in the areas of interest in the published works.

3. TOPIC MODELING

Topic modeling (TM) is an unsupervised machine learning technique used for text mining applications to discover topics that best define a large collection of documents. A "topic" is defined by a collection of words that occur together. Topic models can relate words with similar meanings and distinguish the uses of words with various meanings. Being an unsupervised technique, in TM, documents need not be labelled based on the topics they belong to beforehand. TM is widely used in areas where identifying topics of documents are valuable in a large document collection. One of the algorithms used for TM widely is Latent Dirichlet Allocation (LDA).

LDA algorithm creates topics using word frequencies. The abstracts of the papers are preprocessed to generate document-term matrix which is also called a bag of words. LDA creates topics based on the probability distribution of the words [30][31]. LDA assumes a predefined topic number and discovers topics based on the distribution of words in the documents. It considers each document as a combination of topics. It maps each document in a collection to one of the topics based on the frequency of the words it contains.

4. RESEARCH METHODOLOGY

This study has been conducted as a scoping study to discover research trends in EmpSWE domain. It aims to mine topics and trends in EmpSWE domain with the use of keywords extracted from the collected article in question. The research methodology is shown in Figure 1.

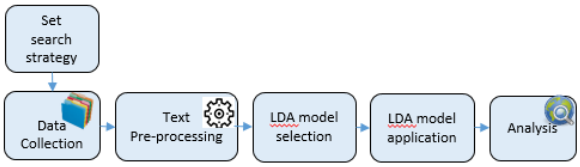


Figure 1. Research Methodology

Hence, answer to the following research questions are explored:

RQ1: What are the most popular research topics and trends?

RQ2: How is the funding distributed between topics?

RQ3: Which are the most productive countries?

To answer the research questions, data were collected from SCOPUS. SCOPUS database was selected to gather articles published in EmpSWE domain between 1970-2019. SCOPUS has an extensive coverage in computer science area as it files articles from various databases like IEEE, ACM, Springer and Elsevier [3].

The search strategy guarantees the replicability of the conducted research. For search strategy, we identified PICOC (Population, Intervention, Comparison, Outcome, Context) terms as suggested by [32] that are derived from the research question. Multi-term search was performed based on the Interventions provided in ref. [11]. Moreover, PICOC terms for keywords selected based on the suggestions of [32]. Additionally, in order to have full coverage of the domain, other research interventions stated for EmpSWE domain in [2] were added as well. We limited the search to Computer Science subject area and publications in English language (Appendix A). As research aims to investigate topics in a large collection of articles published in EmpSWE area, cloud-based computing environment has been selected for text mining process, namely Google's Colaboratory platform with 13GB RAM, 108GB disk capacity and Intel(R) Xeon(R) 2xCPU @ 2.30GHz. that provides a free Jupyter notebook environment for Python implementations.

Real-world databases are extremely prone to noisy, missing, and inconsistent data because of huge collection of data from multiple, heterogeneous sources [33]. In order to get better text mining results, inconsistencies in the data should be removed through preprocessing techniques. Accordingly, several preprocessing techniques have been applied iteratively to increase data quality before the LDA-based topic modeling was applied. Once the data set was collected, cleaning of non-Latin characters, punctuations, special characters and extra spaces have been performed and abstracts were split into words (tokens). Stopwords were removed as well as common terms used for research articles like study, article and research. Lemmatization was applied to get the root form of the words keeping only nouns, adjectives, verbs and adverbs. Additionally, duplications and articles with no abstracts were removed from the dataset. After pre-processing, dictionary and corpus were created for topic modeling.

The corpus created holds the word ids and their frequencies to be used in LDA model. The dictionary created for this research has 10658 documents and 25661 unique words. LDA model was used to discover topics in the collected articles. LDA model implemented in Python's Gensim package and Mallet LDA models were applied and their performances were compared for different number of topics based on the coherence score. Topic coherence score was used to measure how well the topics are extracted in the LDA model. Mallet is a java-based library that implements topic modeling and Mallet topic modeling toolkit contains fast and scalable implementation of LDA algorithm. When two LDA models were compared based on the coherence scores, LDA mallet implementation resulted in higher coherence scores. Hence, in this study, LDA mallet algorithm was used for identifying the topics on the articles collected on EmpSWE research from SCOPUS.

5. RESULTS

In LDA topic modelling, a topic is a group of dominant keywords that are typical representatives of a topic and each article is treated as a collection of topics in a certain proportion. Likewise, each topic is considered as a group of keywords, in a certain proportion. LDA arranges the topic distribution within each article and keyword distribution within each topic to find an enhanced topic-keywords representatives.

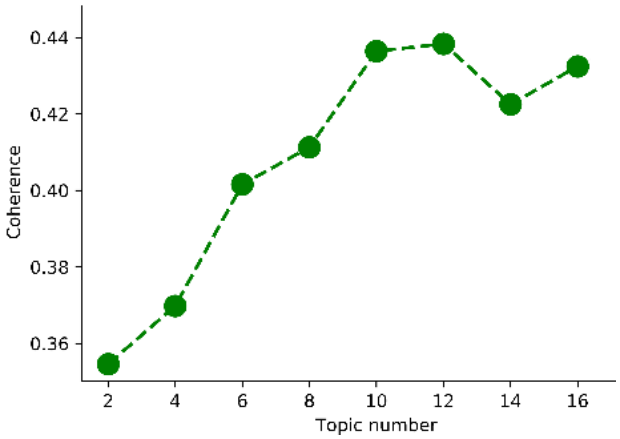


Figure 2. Topic number vs. coherence values

Table 1. Topic-Keyword Distributions

Topic	Keywords
0	0.114*"model" + 0.041*"approach" + 0.033 *"process" + 0.026*"case" + ' '0.025*"base"
1	0.127*"software" + 0.084*"development" + 0.065*"project" + 0.061*"process"+ 0.034*"product"
2	0.037*"problem" + 0.037*"time" + 0.029* "performance" + 0.026*"algorithm" + 0.016*"experiment"
3	0.051*"research" + 0.022*"software" + 0.022*"practice" + 0.021*"identify" + 0.018*"context"
4	0.160*"system" + 0.078*"design" + 0.026*"control" + 0.024*"component" + 0.022*"architecture"
5	0.072*"software" + 0.033*"change" + 0.026*"code" + 0.022*"developer" + 0.015*"pattern"
6	0.051*"service" + 0.043*"application" + 0.034*"network" + 0.024*"web" + 0.018*"technology"
7	0.049*"method" + 0.042*"test" + 0.033*"technique" + 0.030*"base" + 0.022*"approach"
8	0.033*"student" + 0.022*"learn" + 0.020* "computer" + 0.019*"group" + 0.016* "experience"
9	0.025*"experiment" + 0.023*"simulation" + 0.014*"method" + 0.011*"analysis"+ 0.010*"image"
10	0.073*"datum" + 0.058*"information" + 0.041*"user" + 0.041*"knowledge" + 0.016*"support"
11	0.033*"factor" + 0.018*"company" + 0.016*"organization" + 0.015* "relationship" + 0.013*"risk"

The topics dominating EmpSWE field were identified through various trials using different topic numbers. For this purpose, topic number (k) is varied over different LDA models. For each k, a LDA model was created and coherence parameters were calculated as given in Figure. 2.

Coherence values calculated for each topic number (k) and the model that results in the highest value of coherence was chosen as 12. For 12 topics, the top 5 most frequently identified keywords are listed in Table 1. The weights in each topic reflect how important a keyword is to that topic.

By analyzing the most frequent 5 terms representing each topic, topic names were inferred, as given in Table 2.

EmpSWE studies usually involve human aspect within or without the organization aspect [34], hence two of the identified topics revealed this fact. The term empirical software engineering is defined as "is concerned with the scientific measurement, both quantitative and qualitative, of software engineering process and product"[5], which focuses on three aspects of EmpSWE, namely, product, process and measurement. In this regard, the outcome of conducted LDA modelling can be mapped to these aspects as seen in the Figure 3 below.

Table 2. Topic Names

Topic	Name
0	Model
1	Software Development Process
2	Performance
3	Software Practice
4	System Design
5	Coding
6	Service
7	Testing
8	Human
9	Experiment
10	Data
11	Organization

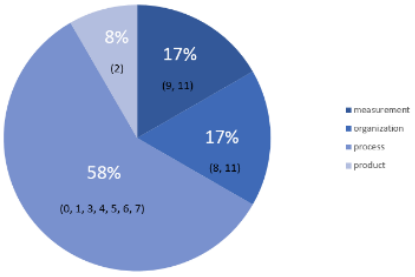


Figure 3. EmpSWE Aspects vs. Topics

The volume and distribution of topics over the collected articles were considered in order to judge how extensively the topics were discussed which is shown in Table 3.

Table 1. Total Number of Articles (1970-2019)

Topic Name	# of Articles	%of Articles
Organization	1226	0.12
Software Development Process	1168	0.11
Human	1077	0.1
Service	912	0.09
Model	893	0.08
Experiment	877	0.08
Performance	848	0.08
Testing	811	0.08
System Design	736	0.07
Coding	720	0.07
Software Practice	706	0.07
Data	684	0.06

Through the years, parallel to the technological developments and their applications in various business domains, we see an increase of interest among scholars in EmpSWE research, especially after 2000's based on the total number of articles (Figure 4).

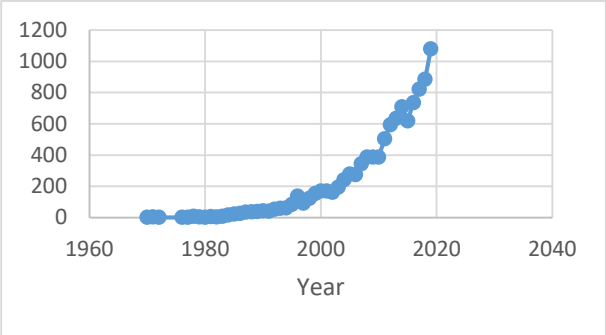


Figure 4. Total number of articles published between 1970-2019

Figure 5 depicts the number of studies in EmpSWE domain conducted in the last 10 years period for each topic. There is an increasing interest in this domain throughout the years especially in the last 10 year period.

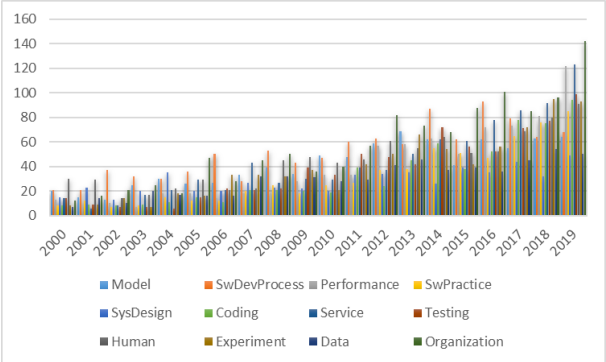


Figure 5. Total number of articles published between 2000-2019 in each topic

For a better understanding of research interests on specific topics, changes in the number of published papers in previous and last decades were compared (Figure 6). As the Figure suggests, there is a giant leap in the number of publications in topics namely, Performance, Coding, Testing and Software Practice whereas the lowest change in interest occurs in System Design.

As seen from Table 3, top three extensively studied topics are Software Development Process, Human and Organization. The change in number of articles specific to these topics are examined further in Figure. 7.

Overall, Organization topic is the mostly studied topic in EmpSWE domain which is followed by Software development process.

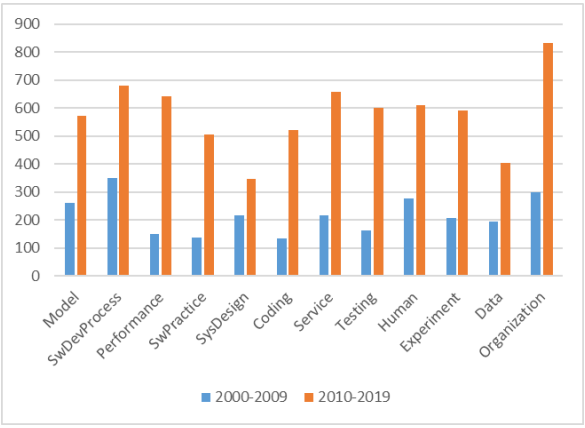


Figure 6. Comparison of number of articles between 2000-2009 and 2010-2019

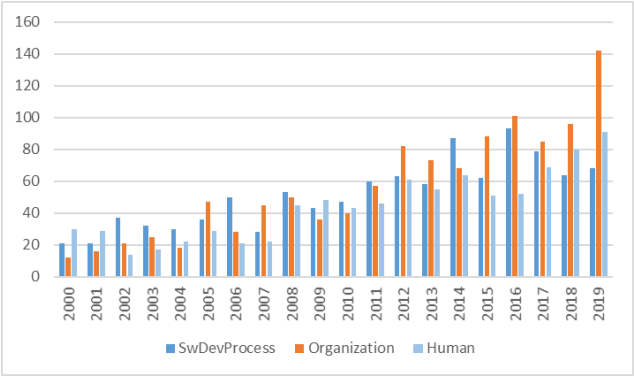
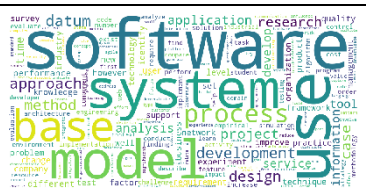


Figure 7. Number of papers published between 2000-2019 in top 3 topics

As EmpSWE aims to shed light to the practical aspects of theories and their applications, such a high frequency of publications on Organization topic is an expected result. 3 most published topics namely Organization, Software Development Process and Human aspects can be considered as the fundamental dimensions of EmpSWE.

Table 4. Word Cloud for Two Decades

1970-1999	
2000-2019	

The word cloud of article abstracts for previous and last decade shows that the research focuses has shifted abruptly (Table 4). In the previous decade, words on model and use become prominent whereas in the last

decade focus shifts to process, design and development aspects. Figure. 8 illustrates how the total number of citations changes in each topic after year 2000. Being mostly published topics, Organization and Software Development Process topics come to the fore. Hence we explored average citations in Figure. 9.

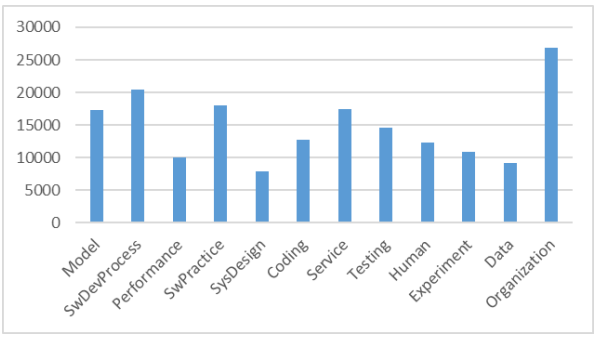


Figure 8. Total number of citations in each topic

As number of studies are high in some of the topic areas, average citation number per topic reveals different insights (Figure 9). When average citation values are considered, Software Practice and Organization topics take the lead.

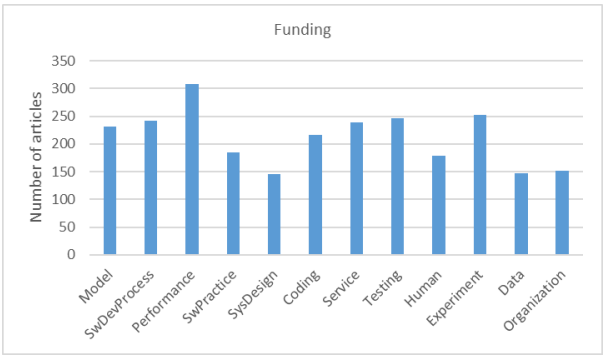


Figure 10. Number of funded studies in each Topic between 2000-2019

When total number of funded studies are considered, in overall, in the collection of 10658 articles only 25% (2694) were funded. 94% of the funded articles were published after year 2000. When topic distribution is considered, Performance stands out as the most funded topic with respect to the total number of funded articles (around 12.1%), whereas System Design and Data topics are the least funded areas (Figure 10). As Performance can be associated with efficiency, cost, time aspects, it is probable that the research in this area receives more interest and gets more funding.

In Figure 13, number of cited and funded articles for each topic is presented. It can be inferred from the results that a research being funded does not necessarily bring about high citation rates.

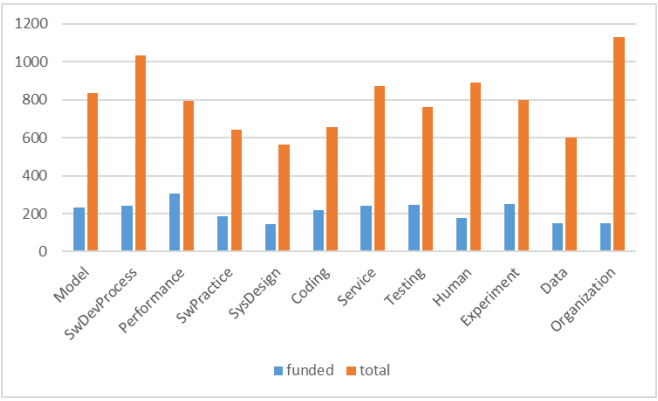


Figure 11. Number of funded vs Total number of articles between 2000-2019

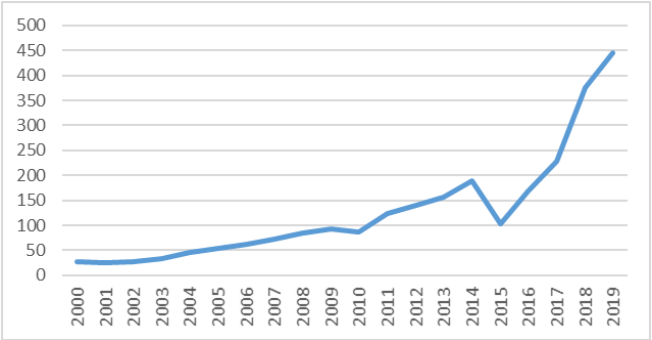


Figure 12. Number of funded articles

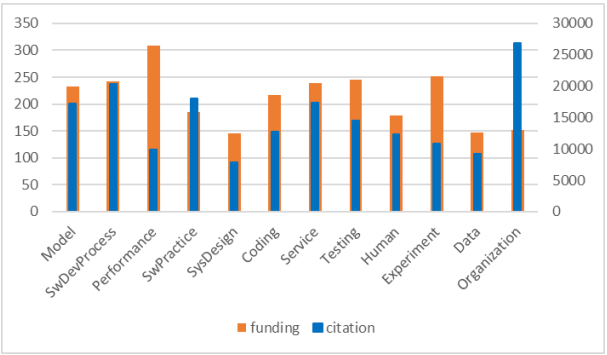


Figure 13. Funding vs Citation of articles

In EmpSWE research, articles got interest from the listed publishers mostly (Table 5). Springer has the highest publication percentage (21%) followed by Elsevier and Kluwer Academic Publishers among the others.

Most productive countries publishing in EmpSWE area was found through author contributions for each article. According to the results, USA, China, United Kingdom, Canada and India are the top 5 countries contributing to research in EmpSWE domain.

Table 5. Top Publishers

Publisher	%
Springer	21
Elsevier	19
Kluwer Academic Publishers	6
Institute of Electrical and Electronics Engineers	4
World Scientific Publishing Co. Pte Ltd	4

6. DISCUSSION AND CONCLUSION

This paper aims to examine the research trends in Empirical Software Engineering (EmpSWE) domain by applying text mining technique, namely LDA based topic modelling. Furthermore, it intends to portray the alterations of research interests within this domain in a comparative manner. The study utilizes SCOPUS database. Based on the collected articles' data from this database, 12 topics were identified with the combination of dominating keywords (Table 1), through LDA topic modelling algorithm. Afterwards, each article was mapped to one of those topics based on the distribution of the keywords in the collected articles' abstracts.

As an overall conclusion, the analysis has revealed that there is an increasing interest to EmpSWE research, especially in the last two decades. The analysis for last 20-year period have shown that; three topics, namely Organization, Software Development Process and Human topics have been studied mostly. As EmpSWE is an area which focuses on practical aspects of software practice [6], it's affect on Organization and Human are expected to be the focus of interest. Moreover, as human resource is the critical asset of the organizations [35], Human and Organization topics have been extensively researched. The least studied topic is Data that constitutes 6% of the total articles. As data science and big data analysis is a growing topic, we expect to have a rapid change in that topic within the coming years.

The fastest growing research areas are found to be in Performance, Coding, Testing and Software Practice topics based on the two decades comparisons (Figure 6), whereas System Design is the slowest growing topic. As the increasing competition between companies upsurges in the digital world [36], quality of the product, efficiency, and performance have become the outmost aspects considered by the companies [37]-[39]. Accordingly, research in EmpSWE area also reveals this fact.

When the number of citations that the articles earned are considered for the last two decades, results of the studies under Software Development Process and Organization topics got the highest interest from the research community which is also expected as these are the mostly published topics (Table 3). However, Software practice and Organization topic are cited most frequently when average citations are considered.

Funding EmpSWE research has been increasing within the last two decades. With the growing competitive pressure in the software industry, funded research in EmpSWE area directed to Performance topic extensively, followed by Coding topic. As Performance can be associated with efficiency, cost, time aspects, it is probable that the research in this area gets more funding from organizations. Similarly, software development effort can be considered as a resource intense, multifaceted issue [40] where coding is the significant part open to new techniques and improvements that require practical assessment.

Countries that have contributed mostly to EmpSWE domain are USA, China and India which are listed in the top 10 list of countries of GDP rankings [41]. Thus, it can be concluded that countries with high economic growth rate are more interested in contributing to this field.

According to the analysis results of topic modelling, it can be expected that the EmpSWE research will gain more interest from the research community as software industry needs for more evidence on the practical applications of the theory. Furthermore, competitiveness pressure of software business will affect research in empirical studies to be directed to more on organizational, human, performance and software development process aspects. Accordingly, as Sjoberg et al. anticipated, there would be more research in the software engineering area with empirical evidence between years 2020-2025, as the trend of last 20 years reveals that EmpSWE research will grow with an increasing speed [8].

As a future study, this research may be extended to gain insights in more specific software engineering research that affects organizational and software development process dimensions like agile and lean practices. Additionally, as grey literature may provide valuable resource for software engineering practice [42], an analysis including practitioner literature for EmpSWE would provide valuable insights and help to develop an advance field, beneficial both for academics and practitioners. Furthermore, comparison of the performances of various topic modeling algorithms may be explored as a future study.

DECLARATION OF ETHICAL STANDARDS

The author of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Gül Tokdemir: All research stages

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Garousi V. and Mäntylä M., "Citations, research topics and active countries in software engineering: A bibliometrics study", ***Computer Science Review***, 19: 56-77, (2016).

[2] Höfer A. and Tichy W. F., "***Status of empirical research in software engineering***", in Basili, V. (Eds.) et al., Empirical Software Engineering Issues, LNCS 4336, 10-19, Springer-Verlag, (2007).

[3] Dieste O., Juristo N. and Martínez M.D., "Software industry experiments: A systematic literature review", ***Ist International Workshop on Conducting Empirical Studies in Industry (CESI)***, 2-8, (2013).

[4] Basili V., "***The Role of Controlled Experiments in Software Engineering Research",*** Empirical Software Engineering Issues, LNCS 4336, 33-37, Springer-Verlag, (2007).

[5] Jeffery R., Scott L., "Has twenty-five years of empirical software engineering made a difference?", ***9th Asia-Pacific Software Engineering Conference***, Australia, 539-546, (2002).

[6] Kitchenham B., "Empirical paradigm - the role of experiments", ***International Conference on Empirical Software Engineering Issues: Critical Assessment and Future Direction***, 25-32, (2006).

[7] Kitchenham B.A., Pfleeger S.L., Pickard L.M., Jones P.W., Hoaglin D.C., El Emam K. and Rosenberg J., "Preliminary guidelines for empirical research in software engineering", ***IEEE Transactions on Software Engineering***, 28(8):721-734, (2002).

[8] Sjoberg D.I.K., Dyba T. and Jorgensen M., "The future of empirical methods in software engineering research", ***Future of Software Engineering***, Minneapolis, MN, 358-378, (2007).

[9] MacDonell S., Shepperd M. and Kitchenham B., "How eeliable are systematic reviews in empirical software engineering?", ***IEEE Transactions on Software Engineering***, 36(5): 676-687, (2010).

[10] Mathew G., Agrawal A. and Menzies T., "Finding trends in software research", ***IEEE Transactions on Software Engineering***, (2018).https://arxiv.org/pdf/1608.08100.pdf .

[11] Dieste O., Grimán A. and Juristo N., "Developing search strategies for detecting relevant experiments", ***Empirical Software Engineering***, 14(5): 513-539, (2009).

[12] Rainer A., "***The Value of Empirical Evidence for Practitioners and Researches***", Empirical Software Engineering Issues, LNCS 4336, 24, (2007).

[13] Malhotra R., "***Empirical Research in Software Engineering, Concepts, Analysis and Applications***", CRC Press, (2016).

[14] Calheiros A. C., Moro S. and Rita P., "Sentiment classification of consumer generated online reviews using topic modeling", ***Journal of Hospitality Marketing and Management***, (2017). http://dx. doi.org /10.1080/19368623.2017.1310075.

[15] Wei X. and Croft W. B., "LDA-based document models for ad-hoc re- trieval", ***in Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval***, 178–185, (2006).

[16] Bauer S., Noulas A., Séaghdha D.O., Clark S. and Mascolo C., "Talking places: Modelling and analyzing linguistic content in foursquare", ***International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Conference on Social Computing (SocialCom)***, (2012).

[17] Vulic I., De Smet W. and Moens M.F., "Identifying word translations from comparable corpora using latent topic models", ***in Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies***, 2, (2011).

[18] Chen Y., Rhaad Rabbani M., Gupta A. and Mohammed Zak J., "Comparative text analytics via topic modeling in banking,", ***IEEE Symposium Series on Computational Intelligence (SSCI)***, (2017).

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

[20] Alghamdi R. and Alfalqi K., "A Survey of Topic modeling in text mining", ***International Journal of Advanced Computer Science and Applications (IJACSA)***, 6(1)(2015).

[21] Albalawi R., Yeap T.H. and Benyoucef M., "Using topic modeling methods for short-text data: A Comparative Analysis", ***Frontiers in Artificial Intelligence***, 3:42, (2020).

[22] Chakkarwar V. and Tamane S. C., "***Quick insight of research literature using topic modeling***", in Smart Trends in Computing and Communications. Smart Innovation, Systems and Technologies, Zhang Y. D., Mandal J., So-In C. and Thakur N. (Eds.), Springer, Singapore, 165: 189–197, (2020).

[23] Vayansky I. and Kumar S.A.P., "A review of topic modeling methods", ***Information Systems***, 94, (2020).

[24] Guzman E. and Maalej W., "How do users like this feature? a fine grained sentiment analysis of app reviews,", ***IEEE 22nd International Requirements Engineering Conference (RE)***, IEEE, 153–162, (2014).

[25] Thomas S. W., Hemmati H., Hassan A. E. and Blostein D., "Static test case prioritization using topic models,", ***Empirical Software Engineering***, 19(1): 182–212, (2014).

[26] Gethers M. and Poshyvanyk D., "Using relational topic models to capture coupling among classes in object-oriented software systems", ***IEEE International Conference on Software Maintenance (ICSM)***, (2010).

[27] Linstead E., Lopes C. and Baldi P., "An application of latent Dirichlet allocation to analyzing software evolution", ***7th International Conference on Machine Learning and Applications***, IEEE, (2008).

[28] Chen T.H., Thomas S.W., Nagappan M. and Hassan A. E., "Explaining software defects using topic models", ***in Proc. of the 9th IEEE Working Conference on Mining Software Repositories***, IEEE Press, 189–198, (2012).

[29] Grimmer J. and Stewart B.M., "Text as data: The promise and pitfalls of automatic content analysis methods for political texts", ***Political Analysis***, 1–31. doi:10.1093/pan/mps028.

[30] Chen H., Xie L., Leung C-C., Lu X., Ma B. and Li H., "Modeling latent topics and temporal distance for story segmentation of broadcast news", ***IEEE/ACM Transactions on Audio, Speech and Language Processing***, 25(1):112–123, (2017).

[31] Bulut A., "TopicMachine: conversion prediction in search advertising using latent topic models", ***IEEE Transactions on Knowledge and Data Engineering***, 26(11), (2014).

[32] Kitchenham B. and Charters S., "***Guidelines for Performing Systematic Literature Reviews in Software Engineering***", Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

[33] Han J., Kamber M. and Pei J., "***Data Mining: Concepts and Techniques***", Morgan Kaufmann Series in Data Management Systems, (2011).

[34] Kampenes V. B., Anda B. and Dybå T., "Flexibility in research designs in empirical software engineering", ***in Proc. of the 12th international conference on Evaluation and Assessment in Software Engineering***, 49–57,(2008).

[35] Joshi M., Sidhu J. and Ubha D., "Reporting intellectual capital in annual reports in Australian software and IT companies", ***Journal of Knowledge Management Practice***, 11(3): 1-18, (2010).

[36] Döring H., "Methodological approaches for research on intangible resources and competitive success in software companies", ***in Proc. of 17th European Conference on Research Methodology for Business and Management Studies***, 123-129, (2018).

[37] Lesser E. and Ban L., "How leading companies practice software development and delivery to achieve a competitive edge", ***Strategy and Leadership***, 4(1): 41-47, (2016).

[38] Garcia F., Cruz-Lemus J. A., Genero M., Calero C., Piattini M. and Serrano M. A., "Empirical studies in software engineering courses: some pedagogical experiences", ***International Journal of Engineering Education***, 24(4), (2008).

[39] Felix A., Huerta R. and Leyva S., "Management of the technological innovation process in software companies from Sinaloa, Mexico", ***Management Dynamics in the Knowledge Economy***, 4(2): 193-214, (2016).

[40] Storer T., "Bridging the chasm: a survey of software engineering practice in scientific programming", A***CM Computing Surveys***, 50(4): 32, (2017). doi:https://doi.org/10.1145/3084225.

[41] http://www.worldpopulationreview.com/countries/countries-by-gdp**,** World Population Review**,** (17.2.2020).

[42] Garousi V., Felderer M. and Mäntylä M.V., "Guidelines for including grey literature and conducting multivocal literature reviews in software engineering", ***Information and Software Technology***, 101-121, (2019).

APPENDIX A

TITLE-ABS-KEY ("SOFTWARE" AND ("EXPERIMENT" OR "EMPIRICAL" OR "EMPIRICAL STUDY" OR "EMPIRICAL EVALUATION " OR "EXPERIMENTATION" OR "EXPERIMENTAL COMPARISON" OR "EXPERIMENTAL ANALYSIS" OR "EXPERIMENTAL EVIDENCE" OR "EXPERIMENTAL SETTING" OR "EMPIRICAL DATA" OR ""SURVEY" OR "CASE STUDY" OR "CORRELATIONAL STUDY" OR "ETHNOGRAPHY" OR "EX POST FACTO STUDY" OR "META ANALYSIS" OR "PHENOMENOLOGY" OR "QUESTIONNAIRE") AND ("INDUSTRY" OR "INDUSTRIES" OR "COMPANY" OR "COMPANIES" OR "BUSINESS" OR "BUSINESSES" OR "ENTERPRISE" OR "ENTERPRISES" OR "INDUSTRIAL" OR "COURSE"))) -- AND (LIMIT-TO (DOCTYPE , "AR") OR LIMIT-TO (DOCTYPE , "RE")) AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (LANGUAGE, "ENGLISH"))