International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 04 - Issue 05 || May 2019 || PP. 99-107

Data Mining Analysis: Research Topics Trend on Web of Science, SINTA, and Student Final Assignment at AMIK Indonesia

Bahruni[1], Fathurrahmad[2]
[1](Department of Information Technology, AMIK Indonesia, Indonesia)
[2](Department of Information Technology, AMIK Indonesia, Indonesia)

Abstract: This study aims to conduct data mining information originating from the Web of Science, SINTA, and Student Final Assignments collected. Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is used as a standard data mining process as well as a research method. The researcher collects data through a Web list of Web of Science, SINTA, and Student Final Assignments. To track trends in research topics by selecting a time span from 2018 to 2019 and exporting data from the Web of Science Core Collection in April 2019. There are 38,162 successful publications taken in the Web-Science-defined category of Computer Science and Information Systems and 230 taken from the SINTA website and the Indonesian AMIK repository are 817. However, I only took 20 of the Highest H-Index Journals on the Web of Science Core Collection. Whereas in SINTA, the author also took 20 Journals with a ranking of SINTA 1 and 2. This research concludes the research topic in the journal Web of Science and is associated with trends in research topics and the ones that emerge most are learning, network, analysis, system, control, data, image, optimization, systems, and neural. The classification uses the Naive Bayes model, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine. Based on the results of accuracy, the Boosted Trees Gradient model has an accuracy of 97.1%, while the Decision Tree is 88.6%, the accuracy value is 88.1% followed by the Naive Bayes model, Generalized Linear Model, Fast Large Margin, Random Forest, and Support Vector Machine. In the Deep Learning model, the accuracy value is 85.2% and the lowest value is in the Logistic Regression model of 42.9%. This shows that the highest level of accuracy is using the Gradient Boosted Trees and Decision Tree models so that the results can predict quite accurately. To author keywords further, the authors found that in keywords that are widely used in the Web of Science Journal by comparing the Indonesian journal SINTA and Repository AMIK, they have a significant degree of difference, but keywords that are not widely used in research on AMIK Indonesia final assignments are keywords. learning while the keyword system has almost the same value. In other words, that of the 10 keywords "system" is still a favorite keyword used in the world of research. The results of the keywords taken can be used as an alternative in determining the final assignment of final students as an improvement in the quality of research based on the latest topics.

Keywords: analysis, trend, data mining, web of science, sinta, final assignment, research reference

I. INTRODUCTION

Professional software products and information technology systems and services are currently being developed by teams and companies that are distributed globally and have become the main success factors in an organization [1]. The development of software is important now and almost touches every corner of our activities, many trends in various countries in the world in 2019 and will be interesting to follow in the following year for software developers and educational institutions.

Data mining is a process of mining information in order to perform classification and prediction in the form of data analysis that can be used to extract models that describe important data classes or to predict future data trends [2]. Data mining is an interdisciplinary field of research that covers several scientific disciplines such as database systems, machine learning, intelligent information systems, statistics, and expert systems [3]. Data mining has developed into an important and active area of research because theoretical challenges and practical applications related to the problem of discovery (or excavation) can be interesting where previously unknown knowledge from the database with real-world facts.

Studies of research topics in publications have gained popularity in the field of science and technology for decades because new trends and topics have emerged quickly and researchers tend to rely on formal channels to communicate research findings [4]. The evolution of research topics in science and technology is of interest to governments, industry, education, and science. Several research techniques have been developed and used in Indonesia [5]. There is a growing interest in research to study the topic of the LIS field, such studies about the evolution of disciplinary frameworks [6,7], longitudinal studies of research subjects [4, 8], journal categorization [9], emerging trends [10, 11], publication patterns [12], and research on topic studies in school librarianship [15], medical librarianship [13], and knowledge management [14]. Although the researchers have

good findings in several fields, research topics for determining the theme of the final assignment of students are still untouched by further investigation. In this study, researchers conducted groupings with keywords and analyze for the last years selected from the Web of Science list, SINTA, and Student Final Assignments.

The author presents a clear description of the research illustrated by comparative topic analysis and longitudinal studies in the final assignment theme category which later journal articles will be selected according to topics related to information systems and computer science. Through this analysis, the authors hope to be able to measure the trend of research topics in the current literature and hope to get further research on trends in the field of information systems and computers.

This research conducted mining using web technology to collect information data originating from Web of Science, SINTA, and Student Final Assignments collected. Thus, by observing the development of the topic of information system management in the world and in Indonesia using data mining can bring a considerable contribution to educational institutions, in an effort to improve curriculum and courses at AMIK Indonesia it is expected that the results of data mining can make reference in determining the final assignment and new courses according to current technological trends.

II. RELATED WORKS

Elvira Asril, FanaWiza and Taslim (2016) conducted research on the application of data mining to explore hidden information in the big data of course values, Mapping was done based on the grades taken by students or prospective graduates, in this case the object of research was students of 2012 s / d 2015 which has reached 120 credits, Then the determination of the subject group is carried out in each of these competencies. The topics studied include databases, data mining, association rules, a priori, and several other possible algorithms, as well as software used for mining processes. Data processing has been prepared using several assistive software such as Excel, and Tanagra. Mining data that has been carried out produces information about the competencies of prospective graduates which can be used as analytical material for decision making [16].

Budiman (2016) conducted research on the quality of human resources of lecturers can be reflected in the productivity and quality of the implementation of tri dharma. The results of the study showed the results of clustering data found a pattern of the proportion of tri dharma into 3 clusters representing patterns: professional lecturers, manager lecturers and lecturers [17].

Assunção, et al (2015) this study discusses approaches and environments to carry out analysis in Clouds for Big data applications. Four important things are done in the field of analytics and Big data, namely (i) data management and architectural support; (ii) model development and assessment; (iii) visualization and interaction users; and (iv) business models. Through detailed surveys, the results of the study identify possible gaps in technology and provide recommendations to the research community in determining research on the future direction of Big data and Cloud computing supported by analytical solutions [18].

Chen, et al (2015) in the journal Data Mining for the Internet of Things: Literature Review and Challenges provides a systematic solution for reviewing data mining in knowledge view, technique view, an application view, including classification, clustering, association analysis, time series analysis and outlier analysis As more devices are connected to IoT, large volumes of data are analyzed, the latest algorithms must be modified to apply to big data. The results of this study produce methods in data mining systems on big data [19].

III. METODE

A. Research Design

Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is used as a standard data mining process as well as a research method such as Figure 1.

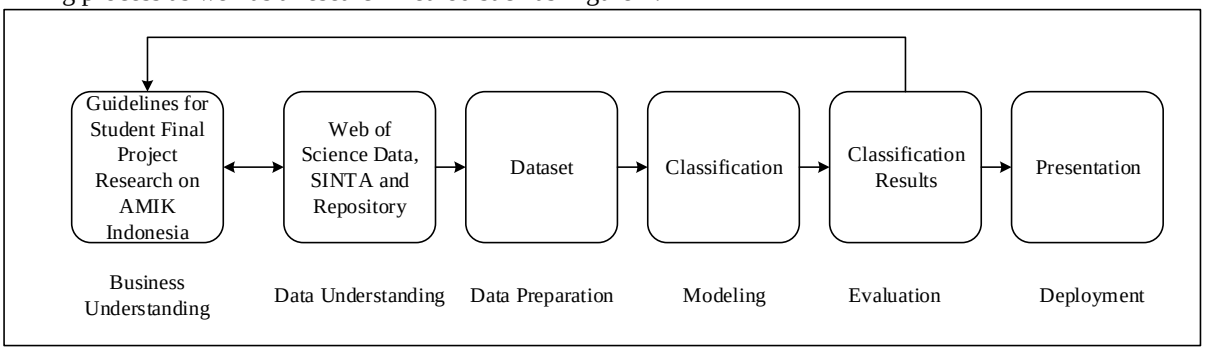

Figure 1. Research Design

B. Research Materials

The researcher collected data through the Web of Science journal, SINTA, and Student Final Assignments and combined several variables to create a journal list with the theme of the final assignment of students, including recording h-index, impact factors, and ranking in SINTA 1 through 2. After searching the Journal title in the database, researchers chose for this research material, as shown in table 1. To track the trend of research topics, researchers chose a time span from 2018 to 2019 and exported data from the Web of Science Core Collection in April 2019.

There were 38,162 publications that were successfully taken in the Web-Science-defined category of Computer Science and Information Systems and 230 were taken from the SINTA website. However, I only took 20 Highest H-Index Journals on the Web of Science Core Collection. Whereas in SINTA, the author also took 20 journals with rankings of SINTA 1 and 2 as shown in table 2. The journal retrieval was taken in view of the fact that journals with ranking SINTA 1 and 2 are nationally indexed and internationally reputed journals (Arjuna, 2019). Whereas in the AMIK Indonesia Repository, researchers took data from 2004 to 2015 involving data on 817 final student studies.

There are two types of keywords available in the dataset, including Author keywords and Keyword Plus. According to Clarivate Guidelines Analytics (2018), Keyword Plus is generated from the title of the article cited by the Web of Science algorithm, intending to add traditional keywords. However, this type of keyword is not suitable for research because the terms generated by the algorithm are also broad to reflect a particular topic. For example, Keyword Plus contains keywords such as Information, Models, Systems, Management, Computers, and so on.

Table 1. Scientific journals selected for this study

No	Journal Name	Indexing
1	Bioinformatics	*Web of Science index*
2	IEEE Transactions on Pattern Analysis and Machine Intelligence	*Web of Science index*
3	IEEE Transactions on Automatic Control	*Web of Science index*
4	IEEE Transactions on Information Theory	*Web of Science index*
5	IEEE Transactions on Image Processing	*Web of Science index*
6	IEEE Transactions on Signal Processing	*Web of Science index*
7	IEEE Transactions on Industrial Electronics	*Web of Science index*
8	IEEE Journal on Selected Areas in Communications	*Web of Science index*
9	Journal of Computational Physics	*Web of Science index*
10	IEEE Communications Magazine	*Web of Science index*
11	MIS Quarterly Management Information Systems	*Web of Science index*
12	IEEE Transactions on Medical Imaging	*Web of Science index*
13	Expert Systems with Applications	*Web of Science index*
14	IEEE Transactions on Wireless Communications	*Web of Science index*
15	BMC Bioinformatics	*Web of Science index*
16	IEEE Transactions on Neural Networks and Learning Systems	*Web of Science index*
17	ACM Transactions on Graphics	*Web of Science index*
18	Pattern Recognition	*Web of Science index*
19	International Journal of Computer Vision	*Web of Science index*
20	IEEE Transactions on Fuzzy Systems	*Web of Science index*
21	International Journal of Electrical and Computer Engineering	SINTA (S1)
22	TELKOMNIKA (Telecommunication Computing Electronics and Control)	SINTA (S1)
23	Indonesian Journal of Electrical Engineering and Computer Science	SINTA (S1)
24	International Journal of Power Electronics and Drive Systems	SINTA (S1)
25	International Journal on Electrical Engineering and Informatics	SINTA (S1)
26	International Journal on Advanced Science, Engineering and Information Technology (IJASEIT)	SINTA (S1)
27	Journal of Engineering and Technological Sciences	SINTA (S1)
28	Journal of ICT Research and Applications	SINTA (S1)
29	Indonesian Journal of Electrical Engineering and Informatics	SINTA (S1)

30	JurnalSistemInformasi (Journal of Information System)	SINTA (S2)
31	JSINBIS (JurnalSistemInformasiBisnis)	SINTA (S2)
32	Register: JurnalIlmiahTeknologiSistemInformasi	SINTA (S2)
33	JurnalTeknologi dan SistemKomputer	SINTA (S2)
34	JurnalTeknologiInformasi dan IlmuKomputer	SINTA (S2)
35	Indonesian Journal of Science and Technology	SINTA (S2)
36	LontarKomputer :JurnalIlmiahTeknologiInformasi	SINTA (S2)
37	KhazanahInformatika: JurnalIlmuKomputer dan Informatika	SINTA (S2)
38	Journal of Degraded and Mining Lands Management	SINTA (S2)
39	Indonesian Mining Journal	SINTA (S2)
40	Jurnal RESTI (RekayasaSistem dan TeknologiInformasi)	SINTA (S2)
41	Student Research at AMIK Indonesia	AMIK Indonesia Repository

Therefore, researchers only take the Author keywords that were adopted in this study. After extracting data, the author can extract 14,099 valid keywords from the Web of Science journal articles, SINTA, and Student Final Assignments. Information and the overall percentage of selected research keywords from the two levels are presented in Figure. 2.

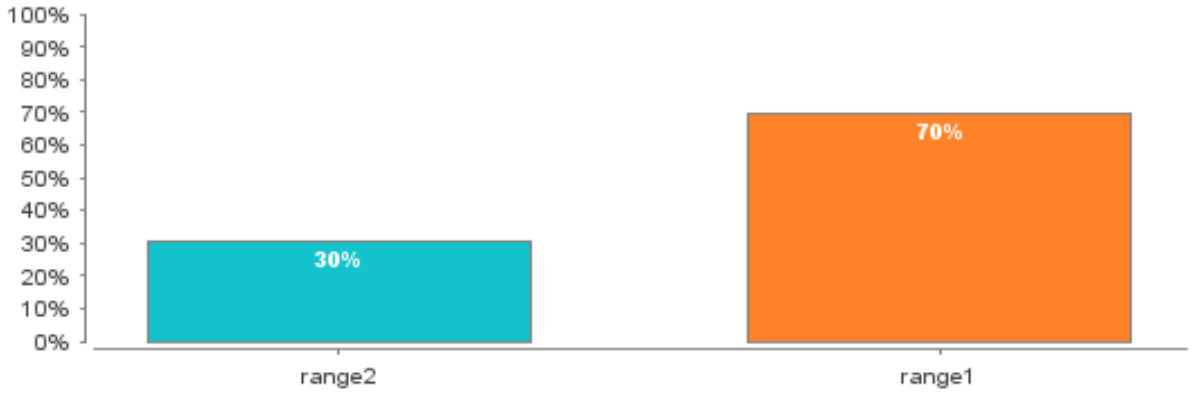

Figure 2. Graph of valid extracted keywords

Keyword grouping and grouping techniques have also been applied in research to streamline the research process and optimize analytical results. The tools or tools used in this study are using Rapid Miner 9.2 Software, as a supporter of data processing using Microsoft Excel 2016. In processing text, there will be stages such as tokenizing, stemming, stop words, and n-grams. The steps for processing text testing will be Process Documents From File, and Sub Process Documents From Files, including Tokenize, Transform Case, and Stop words Filter. The process of applying the model is done to the results before the process which is a data representation in the form of a vector space model. The first method is the application of the Naive Bayes model, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine**.**

C. Research Tools

The tool used in this study is a computer set with standard specifications. Then the software used includes:

1. SQL, database and data mining modeling applications.
2. Ms. Office Excell 2016,
3. Rapidminer 9.2.
4. Windows 8.1 Operating System.

IV. RESULTS AND DISCUSSION

Difficulties in determining the classification of students' final assignment themes are often experienced by each college. The purpose of this study is to provide decision support for policymakers in study programs, especially in AMIK Indonesia so that the title of each student's final assignment is delivered according to the trends of topics contained in Web of Science, SINTA, and Student Final Assignments. The use of the Naive Bayes algorithm, Generalized Linear Model, Logistic Regression, Fast Big Margin, Deep Learning, Decision

Tree, Random Forest, Boosted Trees Gradient, and Support Vector Machine used will prove the level of classification of better trends on research topics.

A. Performance

From the results of cross validation, the results of accuracy and classification of errors obtained from the use of the Naive Bayes model, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine.

B. Accuracy

Accuracy results using the Naive Bayes model, Generalized Linear Model, Logistic Regression, Fast Big Margin, Deep Learning, Decision Tree, Random Forest, Boosted Trees Gradient, and Support Vector Machine as shown in Figure 3 below:

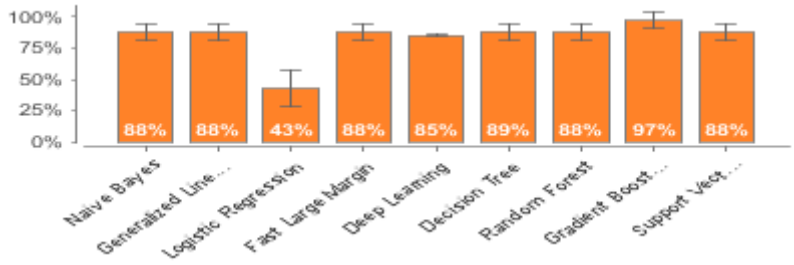

Figure 3. Results of Model Accuracy

In evaluating the accuracy, we can know the accuracy of each model and know the highest accuracy value by comparing the accuracy of the model. Based on the results of accuracy, the Boosted Trees Gradient model has an accuracy of 97.1%, while the Decision Tree is 88.6%, the accuracy value is 88.1% followed by the Naive Bayes model, Generalized Linear Model, Fast Large Margin, Random Forest, and Support Vector Machine. In the Deep Learning model, the accuracy value is 85.2% and the lowest value is in the Logistic Regression model of 42.9%. This shows that the highest level of accuracy is using the Gradient Boosted Trees and Decision Tree models so that the results can predict quite accurately. For the level of classification error in the Naive Bayes model, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine as shown in figure 4 below:

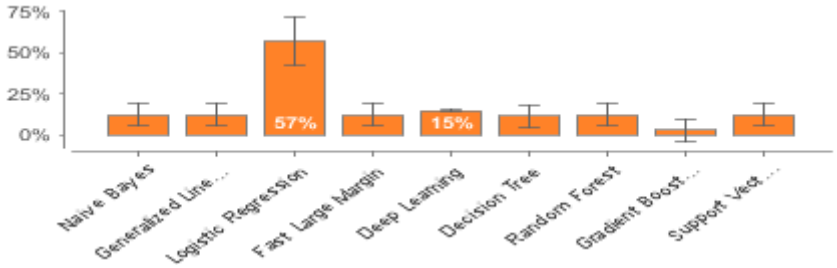

Figure 4. Results of a Classification error

From Figure 5 above it can be seen that the highest level of error classification is in the Logistic Regression model with a percentage of 57.1%, then in the Deep Learning model it has a classification error of 14.8%. in the Naive Bayes model, Generalized Linear Model, Fast Large Margin, Random Forest, Support Vector Machine, and Decision Tree classification error of 11.4%. The lowest classification error was 2.9% in the Gradient Boosted Trees model. The results of the process of using the Boosted Trees Gradient model and decision tree can be seen in Figures 5 and 6.

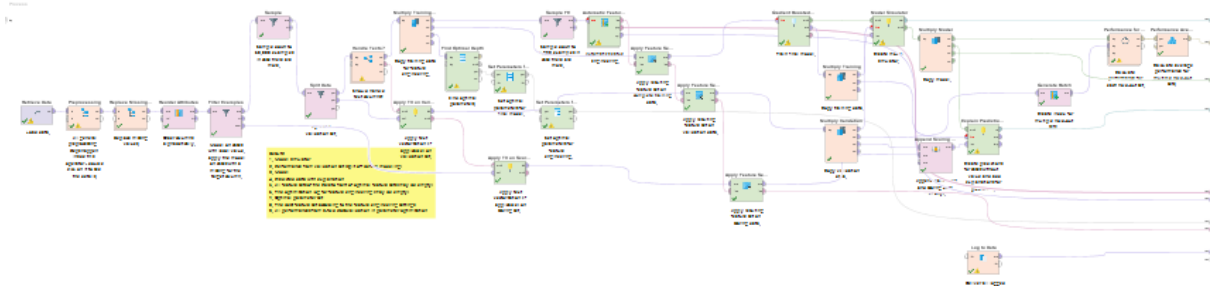

Figure 5. Bossted Trees Gradient Process Model

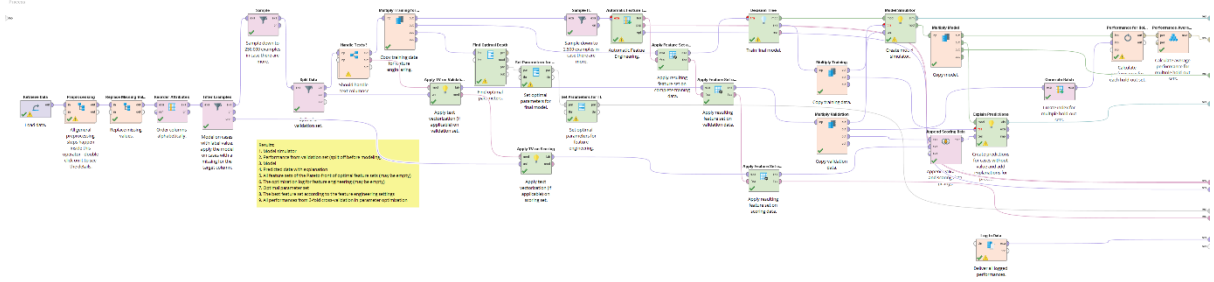

Figure 6. Process Model Decision Tree

C. Discussion

By analyzing author keywords in the Web of Science journal, SINTA, and Student Final Project as shown in table 1, the authors found that the total number of the top ten keywords in the article 2 (two) range journals represented almost 70% and 30% of the total number of words key in this category. Some of the most frequently used terms are even indexed more than 60 times, including 84 times learning, 45 times Network, 41 times Analysis, and 28 times as shown in Figure 7. In Figure 7, the top ten keywords in the past year are described. sourced from the web of science data and SINTA.

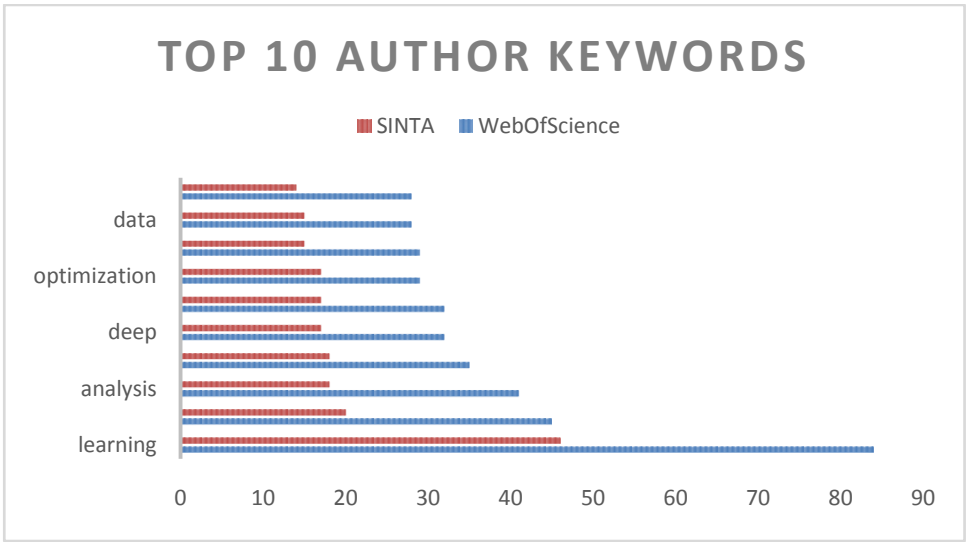

Figure 7. Top Author keywords Web of Science and SINTA

This finding suggests several recent studies. Guoying Liua and Le Yang (2019) found that data is one of the most frequently used keywords. Keyword analysis and systems are also similar to findings by Chang et al. (2015) and Tuomaala et al. (2014). However, this study also revealed several new terms that were first identified as top keywords, including learning, network, analysis, system, control, data, image, optimization, systems, and neural. Moreover, some previous author keyword words can be used lower because of quotes and references [4]. The new findings confirm that when different methodologies are used, especially the selection of journals, the results can be different and consequently more applicable to groups of certain disciplines. This is also because the time period chosen for the study is the latest compared to previous studies. Therefore, researchers in other fields can understand the latest topic trends by choosing popular research [5].

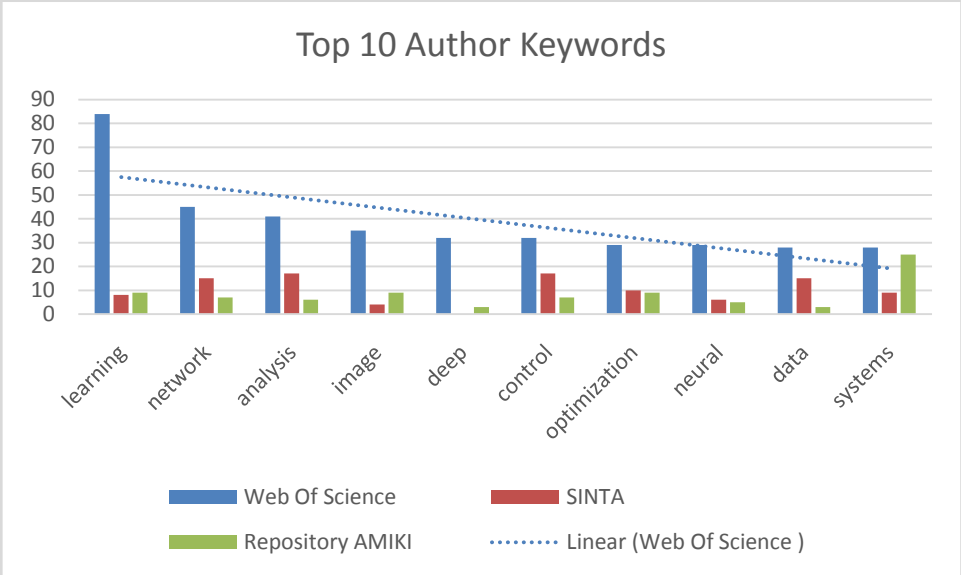

Figure 8. Top 10 Author Web of Science, SINTA and AMIK Indonesia Repository

For further keyword authors, the authors found that in keywords that are widely used in the Web of Science Journal by comparing Indonesian journal SINTA and AMIK Repositories, they have a significant degree of difference, but keywords that are not widely used in AMIK Indonesia final research assignments is a keyword. learning while the keyword system has almost the same value. In other words, that of the 10 keyword systems is still the favorite keyword used in the research world.

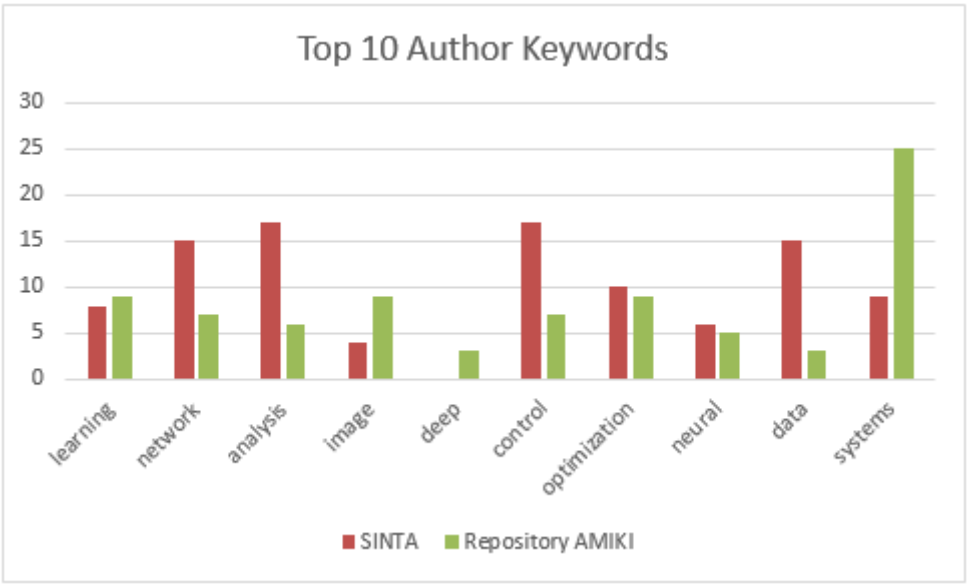

Figure 9. Top Author keywords SINTA and Repository AMIK Indonesia

Looking at author keywords based on selected journal articles SINTA, and Student Final Assignments, the most widely used keywords are the system and are followed by analysis, network, and "data" keywords as shown in figure 9.

V. CONCLUSION

Consider the limitations of previous studies on journal topic research in the field of computer science and systems by using keyword grouping analysis methods on data taken from the Web of Science Core Collection, SINTA, and AMIK Indonesia's reliable repository. Through analysis of journal lists with different levels, this study concludes the research topic in the journal Web of Science and is associated with the trends of research topics and the ones that emerge most are learning, network, analysis, system, control, data, image,

optimization, systems, and neural. The classification uses the Naive Bayes model, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine. Based on the results of accuracy, the Boosted Trees Gradient model has an accuracy of 97.1%, while the Decision Tree is 88.6%, the accuracy value is 88.1% followed by the Naive Bayes model, Generalized Linear Model, Fast Large Margin, Random Forest, and Support Vector Machine. In the Deep Learning model, the accuracy value is 85.2% and the lowest value is in the Logistic Regression model of 42.9%. This shows that the highest level of accuracy is using the Gradient Boosted Trees and Decision Tree models so that the results can predict quite accurately. To author keywords further, the authors found that in keywords that are widely used in the Web of Science Journal by comparing the Indonesian journal SINTA and Repository AMIK, they have a significant degree of difference, but keywords that are not widely used in research on AMIK Indonesia final assignments are keywords. learning while the keyword system has almost the same value. In other words, that of the 10 keyword systems is still the favorite keyword used in the research world. The results of the keywords taken can be used as an alternative in determining the final assignment of final students as an improvement in the quality of research based on the latest topics.

VI. ACKNOWLEDGEMENTS

A big thank you to the Directorate General of Strengthening Research and Development at the Ministry of Research, Technology, and Higher Education as research funders in the 2019 Beginner Lecturer Research scheme. It also did not escape, saying to LPPM AMIK Indonesia who had provided support in this research.

REFERENCES

[1] Ebert, C., Kuhrmann, M. and Prikladnicki, R., 2016, August. Global software engineering: evolution and trends. In Global Software Engineering (ICGSE), 2016 IEEE 11th International Conference on (pp. 144-153). IEEE.

[2] Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

[3] Deogun, J.S., Raghavan, V.V., Sarkar, A. and Sever, H., 1997. Data mining: Trends in research and development. In Rough Sets and Data Mining (pp. 9-45). Springer, Boston, MA.

[4] Chang, Y., Huang, M., & Lin, C. 2015. Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. Scientometrics, 105(3), 2071–2087

[5] Liu, G. and Yang, L., 2019. Popular research topics in the recent journal publications of library and information science. The Journal of Academic Librarianship, 45(3), pp.278-287.

[6] Tuomaala, O., Järvelin, K., & Vakkari, P. 2014. Evolution of library and information science, 1965–2005: Content analysis of journal articles

[7] Walters, W.H. and Wilder, E.I., 2016. Disciplinary, national, and departmental contributions to the literature of library and information science, 2007–2012. Journal of the Association for Information Science and Technology, 67(6), pp.1487-1506.

[8] Onyancha, O.B., 2018. Forty-Five Years of LIS Research Evolution, 1971–2015: An Informetrics Study of the Author-Supplied Keywords. Publishing Research Quarterly, 34(3), pp.456-470.

[9] Abrizah, A., Noorhidawati, A. and Zainab, A.N., 2015. LIS journals categorization in the Journal Citation Report: a stated preference study. Scientometrics, 102(2), pp.1083-1099.

[10] Salim, H.K., Padfield, R., Hansen, S.B., Mohamad, S.E., Yuzir, A., Syayuti, K., Tham, M.H. and Papargyropoulou, E., 2018. Global trends in environmental management system and ISO14001 research. Journal of cleaner production, 170, pp.645-653.

[11] Zhou, B., Bentham, J., Di Cesare, M., Bixby, H., Danaei, G., Cowan, M.J., Paciorek, C.J., Singh, G., Hajifathalian, K., Bennett, J.E. and Taddei, C., 2017. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19· 1 million participants. The Lancet, 389(10064), pp.37-55.

[12] Blecic, D.D., Wiberley Jr, S.E., De Groote, S.L., Cullars, J., Shultz, M. and Chan, V., 2017. Publication patterns of US academic librarians and libraries from 2003 to 2012. College & Research Libraries, 78(4), p.442.

[13] Lu-Yao, G.L., McLerran, D., Wasson, J., Wennberg, J.E., Wasson, J.H., Bubolz, T., Lindsay, C.C., Littenberg, B., Flood, A.B., Chang, C.H. and Mulley, A.G., 1993. An assessment of radical prostatectomy: time trends, geographic variation, and outcomes. Jama, 269(20), pp.2633-2636.

[14] Fteimi, N. and Lehner, F., 2016. Main research topics in knowledge management: A content analysis of ECKM publications. Electronic Journal of Knowledge Management, 14(1).

[15] Joo, S. and Cahill, M., 2018. Exploring Research Topics in the Field of School Librarianship based on Text Mining. School Libraries Worldwide, 24(1).

[16] Asril, E., Wiza, F. and Taslim, T., 2016. Penerapan Data Mining Untuk MenggaliInformasi Tersembunyi Dalam Big data Nilai Mata Kuliah. Digital Zone: Jurnal TeknologiInformasi dan Komunikasi, 7(2), pp.129-134.

[17] Budiman, I., Kom, M., Prahasto, I.T., ASc, M. and Yuli Christiyono, S.T., 2012. Data clustering menggunakan Metodologi crisp-dm untukpengenalan Pola proporsipelaksanaantridharma (Doctoral dissertation, Universitas Diponegoro).

[18] Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A. and Buyya, R., 2015. Big data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing, 79, pp.3-15.

[19] Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A.V. and Rong, X., 2015. Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 11(8), p.431047.