khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Word Cloud of UKSW Lecturer Research Competence Based on Google Scholar Source

SuryasatriyaTrihandaru, Hanna Arini Parhusip*, Bambang Susanto, Carolina Febe Ronicha Putri
Postgraduate Study in Data Science, Faculty of Science and Mathematics
Universitas Kristen Satya Wacana
Salatiga
*hanna.parhusip@uksw.edu

Abstract-There is a need in the Universitas Kristen Satya Wacana (UKSW) to identify the research competence of their faculties at a study program and University level. To accomplish this requirement, we need to automate the analysis of research output and publications quickly. Research articles are scattered in many publisher systems and journals which may be reputable, unreputable, accredited, and unaccredited. We devise a computer code to quickly and efficiently retrieve publication titles recorded in Google Scholar using a machine learning algorithm. The result display is in the form of a word cloud so that dominant and frequent words will be prominent in the visualization. In determining scientific terms to display, we used a modified version of the word cloud Python module and unmodified Term Frequency - Inverse Document Frequency (TF-IDF) library. The algorithm was tested on publication titles of our study program in UKSW and confirmed directly. The system features the ability to produce a word cloud visualization for an individual faculty, for faculties in a study program, or in the University as a whole. We have not differentiated publication sources, whether they are reputable or unreputable, which might affect the accuracy of competence identification.

Keyword*:* machine learning, word cloud, corpus, research competence

***Article info:** submitted December 31, 2020, revised April 4, 2021, accepted April 19, 2021*

1. Introductions

Efforts to collect data on research results at a university often experience difficulties because the data is not documented in an integrated manner. Universitas Kristen Satya Wacana (UKSW) observed this situation. In addition, to trace the research competence of lecturers, university management cannot automatically search for documents to be able to make global conclusions where the documents are scattered in several groups known as reputable (international) and national journals, accredited and non-accredited journals. In addition, data changes over time, too, cannot be followed quickly. Therefore, a technique is needed to be able to automate reading and classifying lecturer research documents.

One of the used techniques to automate digital documents in research is to classify text [1]. In the literature, classification techniques are compared to study the accuracy of several classification techniques in data classification in the form of text but it is still not easy to read visually. For this reason, in this study, the Word cloud was used to be able to provide classification results more easily where this method has also been done by other authors in analyzing text using Latent Dirichlet allocation [2].

Word cloud is a machine learning algorithm. Machine Learning (ML) has been used in various applications in big data processing such as in the manufacture of artificial intelligence, processing COVID-19 data on the death rate in South Korea [3], diabetes data processing [4], and various other applications. Machine Learning (ML) is translated as machine learning in this article as part of data collection, for example, the study of large amounts of text analysis (big data) [5], pattern recognition, and computational theory in artificial intelligence for the initial process of analyzing skin images and the selection of Melanoma features by staining extraction [6] which builds on the theory, method and application of domains related to big data [7]. Likewise, ML is also often related to data mining where data mining explores text data analysis [8]. As a preliminary research, the Word Cloud algorithm was studied to collect spam and non-spam emails [9]. The machine will remember how based on previous knowledge that the email was said to be spam by the user, the next incoming email could be classified as spam and not spam. Such a working approach is called 'learning by reminder'.

This has a deficiency in the learning aspect, namely the ability to label invisible e-mail messages. A learning is successful if it can make progress individually in making wider links. To reach the link in the task of filtering emails on spam, students can search for previously viewed emails and extract words in messages that are indicated as spam. When a new email arrives, the engine can test whether the words in the email are spam and suspect the label [9]. Such a system can correctly predict labeling in invisible emails.

Based on this knowledge, the Universitas Kristen Satya Wacana (UKSW) lecturer research data was then classified, which is expected to be done automatically. The urgency of this research is shown by the need for universities to identify the competence of lecturers in each study program, each faculty and demonstrate the university's excellence through regularly documented research data that can be reported easily where in November 2020 the leadership needs the results of this research immediately. By using the classification techniques that have been studied for spam and non-spam e-mails, this article explains the results of research on UKSW lecturers' research data based on google scholar data, in which more than 2 classifications are formed based on the Word cloud. Lecturers' research data are classified into study program groups and faculties where all article title information in reputable and unreputable journals as well as accredited and unaccredited is not separated.

2. Method

a. Word cloud creation stage

In the literature, there are several Word cloud generators. However, the existing algorithms need to be adapted to the needs of this study. In principle, the font size of a word in the Word cloud is determined by the frequency with which it occurs. For lower frequencies, the font size can be used immediately. Say the initial size is s_0. For larger frequency values, the letters are scaled, normalized linear. Let the value of t_i be the *i-th* count, t_{max} is the maximum count, while t_{min} is the minimum count, then the font size is scaled in the formulation [10]which gains increasing attention and more application opportunities as the big data time approaches. Currently, there has been some online word cloud generators available for users with simple requests, such as repeating the exact phrase, or collecting the text data from a web page. Moreover, most current word cloud generators cannot support characters other than English, which are limited in English-speaking users. There are also packages for programming languages (such as Python and R :

$$s_i = \begin{cases} \left[\dfrac{f_{max}(t_i - t_{\min})}{(t_{max} - t_{min})}\right], t_i > t_{min} \\ 1 \quad , \quad \text{otherwise.} \end{cases}$$

However, this formula has been formulated in the Word cloud generators in Python so that the author doesn't need to do the formulation. The used algorithm is the NLP (Natural Language Process) algorithm, which is a branch of artificial intelligence that is focused on enabling computers to understand and interpret human language. The NLP process is demonstrated at the following stages.

1) Input data
At this stage, the data are inputted by adjusting the referred system,i.e. data from all UKSW lecturers, whether they have Google Scholar ID or not. The titles data from all articles in the google scholar of each lecturer until July 2020 are documented with the help of the Python program so that all titles are collected automatically.

2) Preprocessing data
The first collected data were titles of research articles that were detected on google scholar from UKSW lecturers who have google-scholar IDs. In this step, we transform these data into a recognizable format for the NLP model. Data are usually incomplete, inconsistent, and/or lacking something or trend and also contain errors. Among them, the data contain several incorrectly written words.

a) Tokenization Steps
The tokenization step is the process of cutting words/text into meaningful words, phrases, or elements called tokens. The steps in tokenization in English are shown in the literature [11] where in this study the tokenization for word lists in English and Indonesian. The token list is then used as a further process. The nltk library has tokenized words to make lists of sentences to be split into words or sentences.

b) Lemmatization Step
The Lemmatization step or Word Stemming has the same objective as the above process, namely reducing the inflected forms of each word to the basic form or root words whose relevance is appropriate to the user or reader [12]. For example: the word eating becomes eat. Lemmatization is close to stemming: the difference is that the stemmer operates a word without any knowledge of the context and therefore cannot distinguish words that have different meanings (e.g. bread eaten and eaten bread are clearly different, but here both are not distinguished). Stemmers are usually easier to implement and faster and the drop inaccuracy can be insignificant for some applications. However, because the data are in the titles of articles, this lemmatization was not carried out in this study because the titles were considered singular or unique.

The following are the complete steps that make up the preprocessing stages:

(a) Removes blank rows from the data if any
(b) Changes all letters in lowercase

(c) Token words
(d) Remove the stop word

As mentioned above, lemmatization was not performed in this case.

3) Prepare training data and test data
As in the Machine learning step, we need to separate the data into training data and test data. Training data are taken from some data that have been prepared where direct communication has obtained information on the characteristics of each lecturer or each study program from the existing data. The training data are carried out by the Word cloud in accordance with the information provided.

b. Example of classification testing with the Word cloud for spam and non-spam e-mails

As the first step in research, it is necessary to study word cloud on spam and non-spam email data obtained from the internet where the problem of processing text data for email is also an issue that researchers often study [9]. For that reason, you need the Word cloud generator so you need to install Word cloud. If on anaconda, we can install it on the anaconda menu, namely: pip install wordcloud. Furthermore, 4 DataFrames are created, namely:

a. DataFrame token_vek, contains tokens for each email that has been cleared, in a table of the number of tokens per email. This becomes the feature of the given data.
b. Class_vek dataFrame, contains spam or non-spam classification vectors.
c. Email_vek DataFrame, contains email tokens that have been strung together, for each email.
d. DataFrame vokab_vek, contains the vocabularies used to construct the DataFrame token_vek table.

The word cloud results provide layout, size, and color settings where the word with the greatest frequency will appear the largest and the color is more dominant so that it is easily recognized.

Figure 1 is a Word cloud on the classification of emails in spam. Whereas Figure 2 is a Word cloud in the classification of email which is classified as non-spam. Meanwhile, Figure 3 combines the results of spam and non-spam emails.

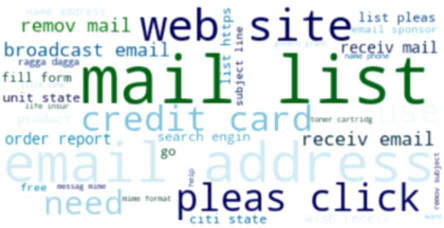

Figure 1. Word cloud results for spam classified emails with data retrieved from the internet https://spamassassin.apache.org/.

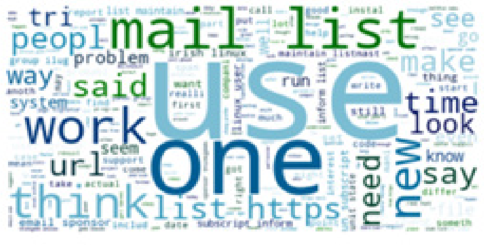

Figure 2. Word cloud results for email classified as non-spam with data taken from the internet https://spamassassin.apache.org/

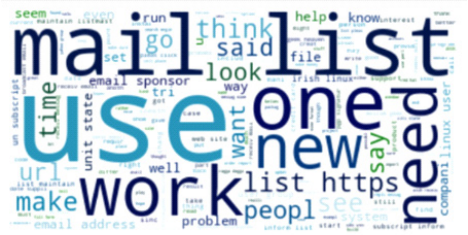

Figure 3. Word cloud results for email classified as spam and non-spam with data retrieved from the internet https://spamassassin.apache.org/.

c. Corpus used

Corpus studies the use of real-life language based on text, describing quantitative and qualitative text analysis techniques. In processing Indonesian text data, programs in the R language can directly accommodate changes from English to Indonesian for the case of text data [13]. In the Word Cloud algorithm that is already in Python, there is already Corpus which is commonly used in text analysis. At first, using the corpus which is shown in the scholarly function. However, this Corpus needs to be modified in this study by defining the Stop-word in this study. This is because the used data are the titles of UKSW lecturer articles that are spread on Google Scholar, from reputable, unreputable journals, accredited and unaccredited journals. So Corpus was built from these research titles indicated on Google Scholar. Therefore, Stop-word is also made in-house where the program is tested several times to improve Stop-word. The words that are considered as Stop-words are shown below.

Case 1.

Words that often appear in various lecturers, study programs, or faculties. The words ‘making’, ‘study’, ‘planning’, ‘learning’, ‘method’ are words that are not unique or special so they need to be discarded.

Case 2.

The word that is mistyped is also a word that needs to be listed in the stop word. For example “studen” is found in the data it needs to be a list of stop words.

Case 3.

Characters such as $,#,@, :, ; , are non-unique words that need to be removed.

It should be noted that the data contain journal titles both in Indonesian and in English. To identify the uniqueness of each lecturer or in the study program group or faculty, words that often appear are certainly not unique. Therefore, the corpus is made separately according to the needs of this research where the data of lecturers must have ID-scholars. Lecturers who do not have ID-scholar cannot contribute to this research.

d. Word Vectorization with TF-IDF

This step is a general process that converts a collection of text documents into feature vectors. There are many methods for this, but the popular one is called TF-IDF ("Term Frequency — Inverse Document" Frequency) which assigns a score to each word. TF-IDF is used directly from the word cloud generator in Python where in general TF-IDF is to summarize how often a word appears in a document (TF) and provide scaling (in descending order) to words that appear in a document (IDF) [14] . So the TF-IDF formulation is not defined by the researcher but directly uses the word cloud generator in Python which contains the TF-IDF formulation. One of the uses of TF-IDF in document classification is shown by a researcher in classifying documents in the types of politics, agriculture, economy, performance, science, and technology [15]. After the corpus is formed as described above, the TF-IDF builds a collection of words that have been learned from the corpus data and which designates a single integer to these words.

3. Result and Discussion

As mentioned in the preprocessing process, the use data are data on titles of articles from UKSW lecturers' research results from Google Scholar. For that, the researchers first recorded all the data of UKSW lecturers, for those with Google Scholar IDs and those without IDs. This process leads to obtaining 390 lecturers with ID-google scholar. With the Python program following the method above, all the titles of articles that have been documented by Google Scholar can be detected by the program. To test the program for this data, The Word cloud was conducted for the Master of Data Science lecturers where the author of this study was also part of the data.

a. Case data Word cloud lecturer Master of Data Science

As a studied initial case, the research data are from the Data Science Master Program lecturers at the UKSW Science and Mathematics Faculty (FSM). By using the method described in Chapter 2, the results of the Word cloud are obtained according to Figure 4-8. As in the explanation of the method, the TF-IDF is not explicitly defined by the author but has been defined directly in the word cloud generator in python. In this initial study, the results of the word cloud were obtained and then communicated directly to the author on the appearance of the word cloud obtained so that TF-IDF was not carried out further studies considering time constraints and the need for immediate word cloud results for each lecturer as well as each study program and faculty must be reported immediately in November 2020. Therefore, accuracy testing is carried out by directly confirming several lecturers regarding the data in the word cloud. In this case, preliminary research was conducted for 5 Data Science Masters lecturers at the UKSW FSM. The obtained information shows that it is appropriate to the research results. Therefore the research is continued by using data for all UKSW lecturers.

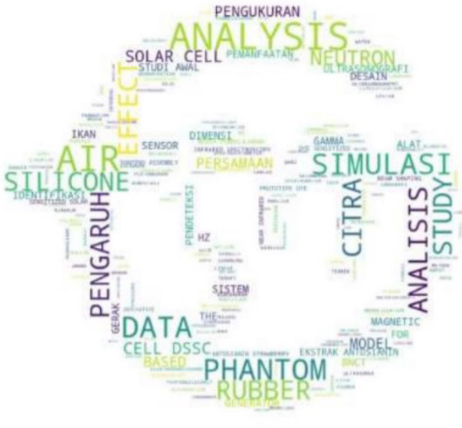

Figure 4. Word cloud for research data, Dr. Suryasatriya Trihandaru, S.Si., MSc.nat based on data up to July 2020.

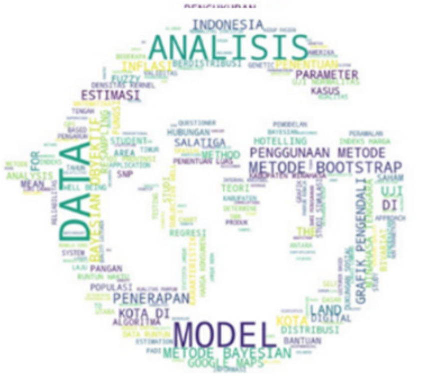

Figure 5. Word cloud for research data, Dr. Adi Setiawan, MS-based on data up to July 2020.

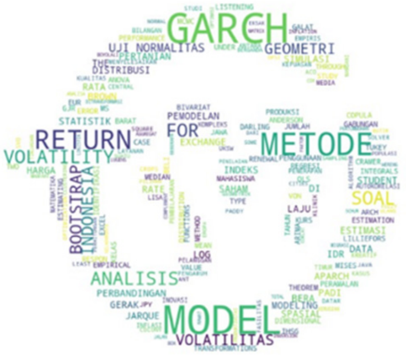

Figure 6. Word cloud for research data, Didit Budi Nugroho, MSi,DSc based on data up to July 2020.

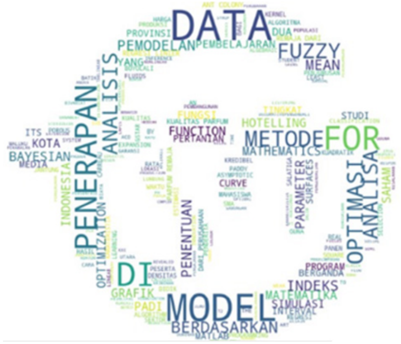

Figure 7. Word cloud for research data, Dr. Hanna Arini Parhusip based on data up to July 2020.

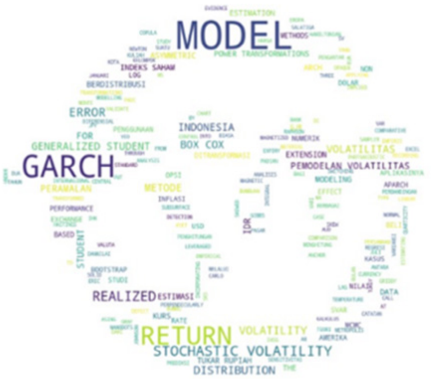

Figure 8. Word cloud for research data, Dr. Bambang Susanto, MS based on data up to July 2020.

b. How to do the analysis?

So far, the Word cloud has not been studied for the accuracy of the model obtained. However, at a glance, the words that stand out from the 5 samples are the word “Model” and the word “Data”. Another dominant word is ‘Analysis’. Broadly speaking, it can be concluded that the research shown by the 5 samples in Figure 4-8 above is about data modeling and analysis. One word that stands out is Garch. From the journal status that appears, it is known that articles with the word “Garch” are included in Scopus, thus gaining dominance in the Word cloud. The next case study will search UKSW’s leading research using the Word cloud for all UKSW lecturers through this research.

c. UKSW Word cloud case data research results

The search for UKSW leading research has been carried out using the Word cloud data for all UKSW lecturers in each of the UKSW faculties. By using machine learning according to the method in Chapter 2, the results of the Word cloud in several faculties are obtained as shown in Figure 9-15.

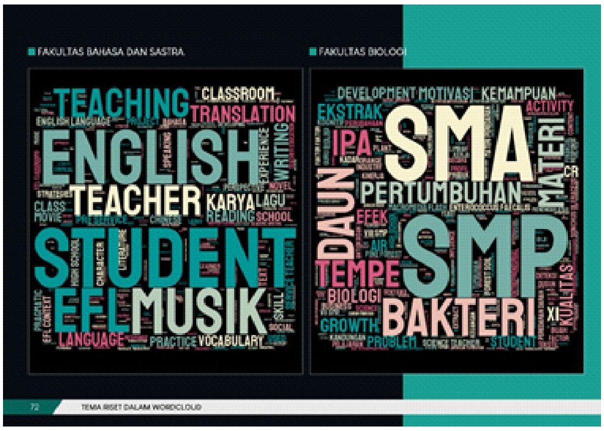

Figure 9. Word cloud for research data from the Faculty of Languages and Letters (left) and the Faculty of Biology (right)

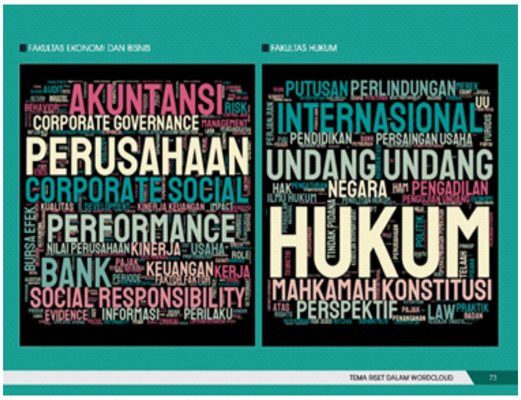

Figure 10. Word cloud for research data from the Faculty of Economics and Business (left) and the Faculty of Law (right) with data up to July 2020.

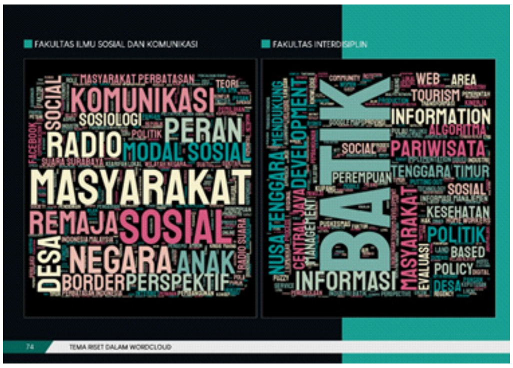

Figure 11. Word cloud for the research data of the Faculty of Social and Communication Sciences (left) and the Faculty of Interdisciplinary (right) with data up to July 2020.

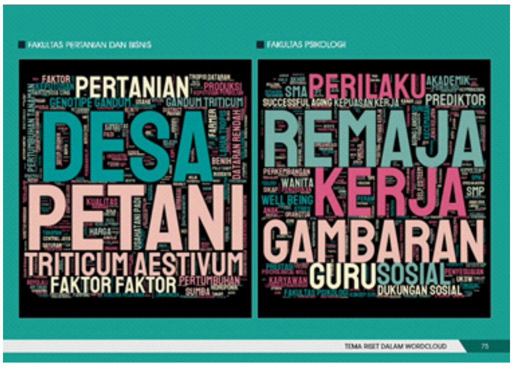

Figure 12. Word cloud for research data from the Faculty of Agriculture and Business (left) and the Faculty of Psychology (right) with data up to July 2020.

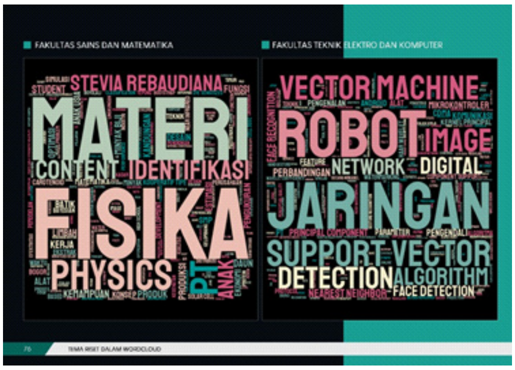

Figure 13. Word cloud for research data from the Faculty of Science and Mathematics (left) and the Faculty of Electrical and Computer Engineering (right) with data up to July 2020.

In the word cloud research that is in UKSW's flagship research, each faculty has different words that stand out. This means that each faculty has different research and results from one faculty to another.

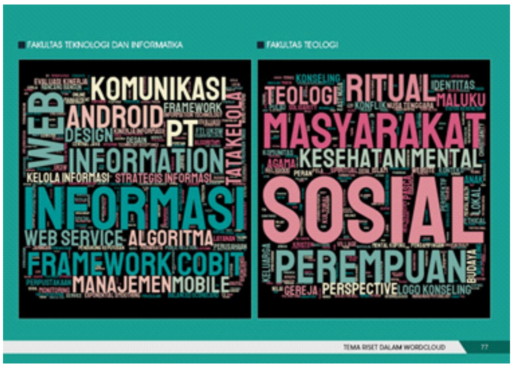

Figure 14. Word cloud for research data from the Faculty of Technology and Informatics (left) and the Faculty of Theology (right) with data up to July 2020.

At the Faculty of Language and Literature, for example, the most prominent words are "English", "Student", "Teacher", "Teaching", and "Music". This means that the Word cloud can show that these words are words with a frequency that stands out from the number of word repetitions in the research at the UKSW Language and Literature Faculty.

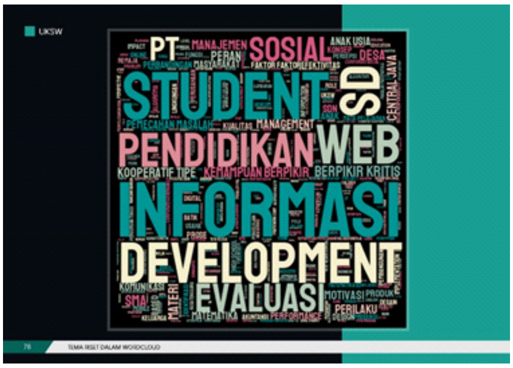

Figure 15. Word cloud for all SWCU research data with data up to July 2020.

It is also shown that the larger the letters, the more popular the topics/reports containing these sentences that are raised in every research in the faculty.

Then in the Faculty of Biology, the words that appear most often are "SMA", "Leaf", "SMP", "Tempe", and "Bacteria". So it can be said that the biology faculty often raises research that uses these words or even research with themes that raise these words such as Tempe or Bacteria. Next in the Faculty of Economics and Business, the words that often appear are "Accounting", "Company", "Corporate Social", "Bank", and "Social Responsibility". This shows that in the Faculty of Economics and Business the word in research or research that appears is around these words. And the most often used are the words Company and Accounting because the words that stand out the most and are the biggest are those two words. At the Faculty of Law, the words that appear most often are "Law", "International", and "law". This shows that in excellent research that is often carried out containing these words, it cannot be separated from the material at the Faculty of Law. In the Faculty of Social Sciences and Communication, the frequency of words that often appear is "Society", "Social", "State", "Communication" and "role" This means that the word frequency that is often used in research or research in this faculty is about those topics. In the Interdisciplinary faculty, the word that often appears is "Batik" because the word that stands out in the word Batik means that in this faculty often uses the word/research containing the word Batik. Then at the Faculty of Agriculture and Business, the frequency of words that often appear is the words "Village", "Farmers", "Agriculture", and "Triticum Aestivum". This shows that the research or research that is often carried out by the Faculty of agriculture and business is to raise the theme of the words that stand out in the Word cloud. Then in the Faculty of Psychology, the words that often appear are "Teenagers", "Work", "Images", "Teachers", and "behavior" this is evidenced by the presence of the most prominent words that have the largest size. This means that in the psychology faculty often raises themes or words that use these sentences. In the Faculty of Science and Mathematics, the frequency of words that often appear are the words “Material”, “Physics”, “Stevia rebaudiana”, and “Identification”. Similarly, this shows that the Faculty of Science and Mathematics often uses these words as research or research being carried out. Then at the Faculty of Electrical and Computer Engineering, the most prominent words are the words “Robot”, “Network”, “Digital”, “Support Vector”, and “Vector Machine”. This shows that the Faculty of Electrical and Computer engineering often uses words or themes on these. Then at the Faculty of Technology and Informatics, the frequency of words that often appear is the words “Information”, “Framework Cobit”, “WEB”, “Communication” and “Android”. Observing the Faculty of Theology, the words “Social”, “Society”, “Ritual”, “Mental Health”, and “Women” appeared most frequently. Finally, the studied is carried out for the whole university. The words that stand out are the words “Student”, “Information”, “Education”, “Development”, “WEB”, and “Evaluation”. This shows that at UKSW, these words are most frequently carried out in the researches done in the Faculty of Theology. The accuracy of the word cloud obtained is confirmed with existing data and communicated directly with the relevant parties. The results are expected as desired. However, it turns out that these six words do not appear dominantly from the titles of reputable articles where reputable journals are considered to have greater weight than those that are not reputable.

The explanation above is done by observing the results of the visual word cloud that appears where the researcher can search quickly because of the prominent appearance of the word in the observed word cloud. This is in accordance with the purpose of the urgency of research, namely to get information quickly on the characteristics of each study program or faculty through the word cloud without paying attention to the TF-IDF that is in each result considering the limited research time and it is necessary to immediately report the research results in November 2020 on related parties so that further management steps can be taken with the results of this research. In addition, other authors who work with word clouds also provide a similar analysis where the analysis is carried out by paying attention to the visual results of the word cloud that are displayed [16]. However, the TF-IDF algorithm can also be used for analysis even with the modified TF-IDF algorithm which weights have been shown to give better results in the case of word analysis of Chinese documents [17]. By comparing these results, it is hoped that in further research, weighting can be carried out on article titles with a higher reputation than journals with lower categories (for example those that are not accredited) so that the results of the word cloud can create a dominant visualization for article titles in reputable articles.

4. Conclusion

In this study, it was demonstrated about making a word cloud for Universitas Kristen Satya Wacana (UKSW) lecturer research data based on the data on lecturer article titles that were documented on Google Scholar. This was done because efforts to collect data on UKSW research results often encountered difficulties because the data were not documented in an integrated manner. In addition, to track lecturers’ research competencies, it cannot be done by tracing documents manually in order to be able to make global conclusions in a timely and continuous manner. The used method is the machine learning method using the word cloud generator in Python where the corpus is built based on the data on the titles of the lecturer articles on Google Scholar so that the corpus in built specifically for this purpose.

The obtained results from this study are the visualization of word cloud leading to find out more easily classification of the dominant topics researches done by lecturers in University in the period until July 2020. These visualizations of word clouds have been carried out based on IDs lecturers of google scholars in study programs, faculties, and at university. Meanwhile, the outstanding results of each lecturer indexed by Scopus or reputable journals have not been identified dominantly on the results of word cloud. Therefore, in further research, this research will be corrected by paying attention to this sorting so that reputable journals get the highest frequency and it appears in the word cloud more dominantly that articles in reputable journals get better ratings than unreputable. Likewise, articles in accredited journals will dominate the word cloud more clearly than articles in unaccredited journals.

Acknowledgement

This research was funded by a UKSW research center with internal research for the 2020/2021 fiscal year with a research entitled “UKSW Leading Research Analysis using Machine Learning and Google Scholar Data”.

Reference

[1] V. B. Kobayashi, S. T. Mol, H. A. Berkers, G. Kismihók, and D. N. Den Hartog, “Text Classification for Organizational Researchers: A Tutorial,” *Organ. Res. Methods*, vol. 21, no. 3, pp. 766–799, 2018, doi: 10.1177/1094428117719322.

[2] R. Kusumaningrum, S. Adhy, and Suryono, “WCLOUDVIZ: Word cloud visualization of Indonesian news articles classification based on Latent dirichlet allocation,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 16, no. 4, pp. 1752–1759, 2018, doi: 10.12928/TELKOMNIKA.v16i4.8194.

[3] C. An, H. Lim, D. W. Kim, J. H. Chang, Y. J. Choi, and S. W. Kim, “Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-

020-75767-2.

[4] J. Beschi Raja, R. Anitha, R. Sujatha, V. Roopa, and S. Sam Peter, “Diabetics prediction using gradient boosted classifier,” *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 3181–3183, 2019, doi: 10.35940/ijeat.A9898.109119.

[5] H. Qian, “Big data Bayesian linear regression and variable selection by normal-inverse-gamma summation,” *Bayesian Anal.*, vol. 13, no. 4, pp. 1007–1031, 2018, doi: 10.1214/17-BA1083.

[6] Y. A. Sari, A. G. Hapsani, S. Adinugroho, L. Hakim, and S. Mutrofin, “Preprocessing of Skin Images and Feature Selection for Early Stage of Melanoma Detection using Color Feature Extraction,” *Int. J. Artif. Intell. Res.*, vol. 4, no. 2, p. 95, 2021, doi: 10.29099/ijair.v4i2.165.

[7] L. Demidova, E. Nikulchev, and Y. Sokolova, “Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 294–312, 2016, doi: 10.14569/ijacsa.2016.070541.

[8] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, “The application of text mining methods in innovation research: current state, evolution patterns, and development priorities,” *R D Manag.*, vol. 50, no. 3, pp. 329–351, 2020, doi: 10.1111/radm.12408.

[9] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems,” *Heliyon*, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[10] Y. Jin, “Development of Word Cloud Generator Software Based on Python,” in *Procedia Engineering*, 2017, vol. 174, pp. 788–792, doi: 10.1016/j.proeng.2017.01.223.

[11] G. Sazandrishvili, “Asset tokenization in plain English,” *J. Corp. Account. Financ.*, vol. 31, no. 2, pp. 68–73, 2020, doi: 10.1002/jcaf.22432.

[12] G. Astika, “Lemmatizing textbook corpus for learner dictionary of basic vocabulary,” *Indones. J. Appl. Linguist.*, vol. 7, no. 3, pp. 630–637, 2018, doi: 10.17509/ijal.v7i3.9813.

[13] Hartanto, “Text Mining Dan Sentimen Analisis Twitter Pada Gerakan Lgbt,” *Intuisi J. Psikol. Ilm.*, vol. 9, no. 1, pp. 18–25, 2017.

[14] N. K. Widyasanti, I. K. G. Darma Putra, and N. K. Dwi Rusjayanthi, “Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia,” *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, p. 119, 2018, doi: 10.24843/jim.2018.v06.i02.p06.

[15] M. Umadevi, “Document Comparison Based on the Page Layout,” no. 1, pp. 2–6, 2020

[16] Y. Huang, Y. Wang, and F. Ye, “A Study of the application of word cloud visualization in college english teaching,” *Int. J. Inf. Educ. Technol.*, vol. 9, no. 2, pp. 119–122, 2019, doi: 10.18178/ijiet.2019.9.2.1185.

[17] J. Chen, C. Chen, and Y. Liang, “Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word,” 2016, vol. 133, pp. 114–117, doi: 10.2991/aiie-16.2016.28.