Stemming Algorithm for the Indonesian Language: A Scientometric View

Aries Maesya
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
aries.maesya@binus.ac.id

Arief Ramadhan
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
arief.ramadhan@binus.edu

Edi Abdurachman
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
edi.abdurachman@binus.edu

Agung Trisetyarso
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
atrisetyarso@binus.edu

Muhammad Zarlis
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
muhammad.zarlis@binus.edu

***Abstract*—Stemming is the process of cutting affixes, both prefixes and suffixes from a term to get the root of the word that has affixes. Stemming can be done in any language, especially in Indonesia Language. Indonesian which is rooted in Malay and Sanskrit has a bigger influence on Language. The algorithm of stemming that has developed is the one that is based on the dictionary of a root word. This kind of algorithm was started by Nazief- Adriani and it grew up to be Confix Stripping (CS), and then was perfected to be Enhanced Confix Stripping (ECS). A scientometric analysis is a mathematical method used to identify academic publications related to citations and scientific matters and is intended for use in libraries or other fields. In this research, the publication of which using stemming algorithm selected and collected using keywords related to Nazief Adriani and ECS. The publications are collected from dimension.ai (an online database for advance scientific research). From those publication we did the analysis bibliometric help by the VOS Viewer application using the analysis technique of co-authorship and citation technique, and all done. The result are, from 310 publications that have keywords Nazief Adriani and 119 publications containing the keyword ECS, it was found that the owning documents will be the reference sources of the next research in which the most citation and co-authorship are found in Indonesia.**

***Keywords*—Stemming, Nazief-Adriani, Enhanced Confix Stripping, Scientometric Analysis, VOS Viewer**

I. Introduction

Stemming is the process of reducing related words to a standard form by removing affixes from them. [1]. Stemming techniques are classified into two categories: basic rules and statistical [2]. Stemming has an important role because stemming results are used to extract features presented in the text. Words that appear in a text have various forms. There is a basic word containing affix. The basic words containing the affix are considered to have different entities. Therefore, in the feature extraction, the two forms will be considered distinctive features. Different features will have different values. This will greatly affect the extraction results of the feature themselves. Due to these reasons, the words containing affixes must be stemmed to determine the basic words, so they will have the same forms. The words are not only prominent but also strengthen their values in the text. (Prihatini et al, 2017). The Stemming process is widely used in information retrieval to improve the quality of obtained information. The quality of the information is mentioned to connect the relationship between one variant word with others. For example, the word 'read' (the act of reading), and "read" (a person who does the activity) originally has different meanings. Those words can be stemmed from the word "read" so the above words are interconnected. Stemming can be used to reduce the size of an index size file. For example, in the description, there are variants of the words "give", "gave", and "given". It has the root of the stem word "give" [3].

Stemming has been applied in European languages, especially English. The algorithms are commonly used by Porter Stemmer to Dawson Stemmer [4]. Different from English, according to Tala in his article entitled: Stemming Effect on Indonesian Retrieval System conducted in 2003 stated that the Indonesian Language has morphological word structures which are special and complex compared to other languages. Commonly, it contains inflection structures and derivatives. Inflection is a collection of suffixes that cannot change the forms and the meanings of basic words. Whereas, Indonesian basic derivative structures can be composed of prefixes, suffixes, or even combinations between them [5]. Moreover, Indonesian stemming is very interesting to be studied and developed because Indonesian is a language that has various morphology affixes. Often a basic word or basic form needs to be given the affixes to be able to be used in the talk. These affixes can change the meaning, the types and the functions of a basic word. Which affixes that should be used depend on the purpose of the user in the talk [6].

Indonesian stemming algorithms which have been more developed are stemming algorithms based on the basic dictionary words. They consist of grouping affixes, word order, and how to remove the affixes as well as match them with the basic word dictionary. The basic word-based stemming Indonesian algorithm has been successfully developed, starting with Nazief Adriani in 1996, then the

Confix Stripping algorithm developed by [7] in 2007, developed into a stemming algorithm called Enhanced Confix stripping (ECS) by [8] in 2009. Further research was conducted by [9] in 2016 by developing algorithms based on the classification of affixes. Further development was carried out by [6] in 2017 which was named the Stemming Algorithm New Enhanced Confix Striping (NECS). In this algorithm, the grouping of affixes is a combination of ECS stemming algorithms and stemming algorithms based on affix classification.

The latest development related to stemming algorithms has been carried out by (Mulyana et al, 2019) in 2019, namely by modifying grouping, sorting, and removing affixes based on Morphophonemics. The development of stemming Algorithms that is based on the Indonesian language can be seen in figure 1.

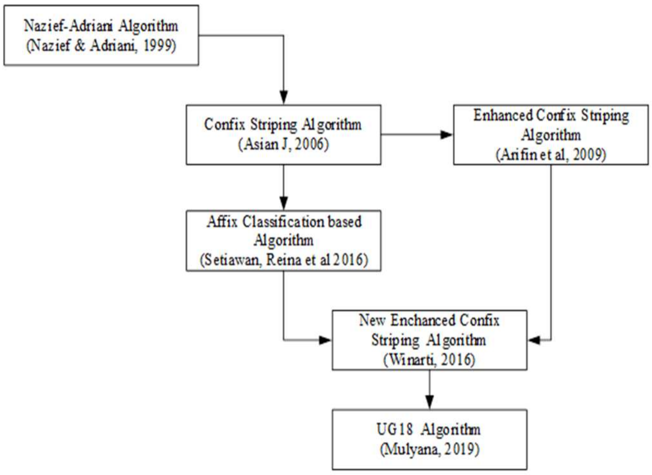


Figure 1. Summary Stemming Algorithm Indonesian Patching Order Patching.

Nazief and Adriani's stemming algorithms were developed by [7] Bobby Nazief and Mirna Adriani in 1996 and became the forerunner of the continuality algorithm developments. The algorithm is based on the rules of the morphology of broad Indonesian which are collected into one group and encapsulated into permissible affixes and impermissible affixes. There is a basic word dictionary used to support word recording and word matching after stemming words [10]. The Nazief Adriani algorithm has been widely used and has good accuracy. This is proved by the research [11] that has been compared to its performance by the previous research Paice-Husk algorithm. In the research, the Nazief-Adriani algorithm produced an accuracy of up to 91.87% with 1799 words successfully stemmed, while the Paice -Husk algorithm with an accuracy of 64.4% and 1261 words were successfully stemmed. The research conducted by [12] can be used as a reference that stemming used by Nazief-Adriani can be combined in completing task analysis sentiment.

Stemming Enhanced Confix Stripping (ECS) algorithm is a development of the confix stripping algorithm. The enhanced confix stripping stemmer algorithm is an algorithm developed by [8] in 2009 to correct the errors of the confix stripping stemmer algorithm which is a development of the stemming algorithm by Nazief and Adriani ECS algorithm makes improvements to stemming errors made by previous algorithm, namely CS [13]. For example, the word "promoting" cannot be stemmed by the CS algorithm [14]. The ECS algorithm has been applied to determine the terms in the grouping of text documents which is a total of 108 articles that have been collected from Semarang State University, ECS has successfully reduced from 199,356 terms to 2,624 terms. The reduced percentage was about 98.68%. The process of searching and indexing translations of the Qur'an tends to be faster after the implementation of ECS stemming [15].

In general, Stemming Enhanced Confix Stripping (ECS) is the same as Nazief and Adriani's stemming algorithms. Both of them complement and improvement in the rules of decapitation [16]. Because Nazief-Adriani has been successfully implemented in various problems and ECS is the development of an algorithm which recently developed, the Nazief-adriani algorithm and Enhanced Confix Stripping were selected in this study to see the influence in the area of \the development of other algorithms and the citation level.

Moreover, scientometric analysis is used as a method to analyze. Scientometrics is a qualitative study based on the level of its publication. This includes identifying the fields of science that appear in the research. In addition, we can examine the research development continuously and geographically based on the research distribution [17]. Scientometrics was introduced by Pritchard, Nalimov, and Mulchencko in1969 [18]. Scientometrics is a study of science existed since 1980 including in the library field of science. But it would gradually be studied and applied throughout the field of science. The scientometric method is a literature measurement-based method using a statistical approach and is an implementation of quantitative analysis to test conventionality and novelty in international research collaborations. It turns out that international collaborations failed to produce more new articles. International collaboration seems to be the result of fewer combinations of new and more conventional knowledge. The costs and barriers to communicating for international collaboration suppress novelty. Higher citation effects can be explained by the audience effect. The more writers are from any country, the greater access results in larger communities [19].

II. RESEARCH METHOD

In conducting scientometric analysis in this research, some steps are taken including the determination of keywords, the process of finding some data, and conducting scientometric analysis.

A. Research Question

There are 3 research questions as follows:

1. What kind of algorithms are widely used for stemming the Indonesian Language?
2. How are the citations of algorithms resulted from RQ1?
3. How are the top most cited algorithms resulted from RQ2 compared to the other algorithms?

B. *Keyword Determination*

The determination of this keyword is the first important step, once we know what problems will be solved and found out for its novelty level. The keywords used in this study are Nazief-Adriani and Enhanced Confix Stripping.

C. *Data Searching*

Dimensions ai is a website used for data searching. This digital science source platform is chosen because it provides a wide range of publications from the 80s to 2022 compared to other platforms that are very limited and restricted. It is easy to be used by typing keywords in the field searching. The process of exporting data is fast and of course free. (Maximum 500 publications which can be exported). Each keyword was taken from publication data from 2013 to 2022. There were 310 publication data for nazief-adriani keywords and 119 publication data for enhanced confix stripping keywords. The publication data is stored by using an exported method with a .cvs extension to be readily analyzed and visualized with the VOSViewer application.

D. *Scientometric Analysis Type*

In this study, 2 types of analysis are conducted as the following:

1. *Co-authorship Analysis*

The technique is used to find out the relationship of various studies based on research documents produced by the researchers. The Co-authorship network is a tool for uncovering the direction of collaboration and identifying researchers and institutions that leading the research [20]. In short, this analysis is used to focus on authorship networks based on the author's name. For instance, the author of cooperation with whom is in his research. The analysis of the units used is based on authors and countries.

2. *Citation Analysis*

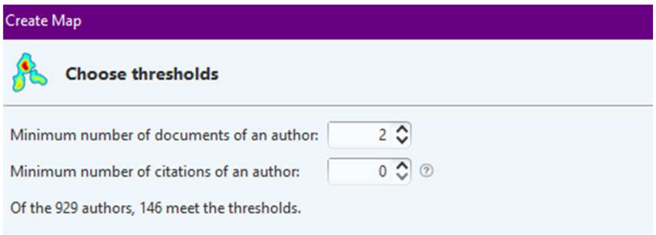
This technique is based on a database of scientific and medical research that can be legally applied to all scientific fields, including applied and technical sciences, social sciences, and humanities, but it is also frequently debated. In science, research groups are natural 'business' units, and therefore it constitutes the most useful level of aggregation in citation analysis because scientific research is a teamwork result [21]. Garfield in his paper argues that the frequency of journal citation is a function both of the published material of scientific of significant and the number of articles publishes each year.

III. RESULT AND DISCUSSION

A. *Analysis Based on Co-Authorship Analysis Technique*

1) *Co-Authorship with Unit Analysis: Author*

In this analysis, the selected unit analysis is based on the authors. The selected counting method is full counting. In the Figure 2, the threshold set is set at number 2 because it will show authors with at least 2 documents in the existing data set. If it has 4 documents, it will not be shown.



Choose thresholds

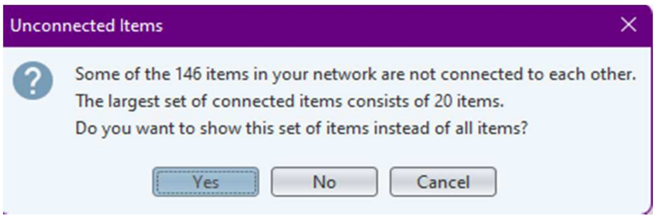
Minimum number of documents of an author: 2

Minimum number of citations of an author: 0

Of the 929 authors, 146 meet the thresholds.

Figure 2: **Thresholds Settings.**

From the regulatory process, 149 authors were selected. Only 20 out of 149 selected authors are linked or connected as shown in Figure 3.



Some of the 146 items in your network are not connected to each other.
The largest set of connected items consists of 20 items.
Do you want to show this set of items instead of all items?

Figure 3: **Successful networked pop-up authors ready to visualize.**

Figure 4 shows that 20 selected authors are divided into 3 large branches, namely 1) green: nazief research, bobby which became the basis of reference for further research, namely [adriani] to [Huda, Fatchul] 2) blue: research belonging to [Fadila, Siti, Dara] which became the basis of reference for research [Lydia, Maya Silvi], [Gunawan], and [Huda, Miftahul], 3) red color: research [Eclipse, Yana Aditia] became a reference for further research, namely [Maylawati] and [Darmalaksana, Wahyudin]. It is shown that Garhana, Yana Aditia has the most co-authors compared to others.

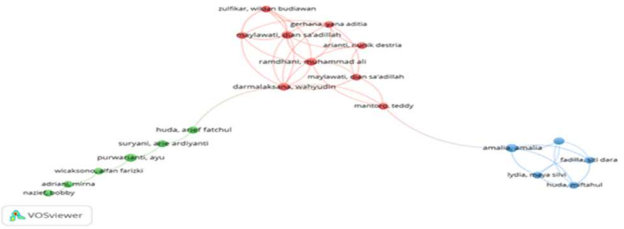


Figure 4: **Co Analysis Network Visualization – Authorship**

The research evolution is demonstrated through the visualization of the visualization overlay feature shown in Figure 5. This evolution is based on years of journals or publications collected. This visualization will be shown in the form of a color bar in the lower right corner.

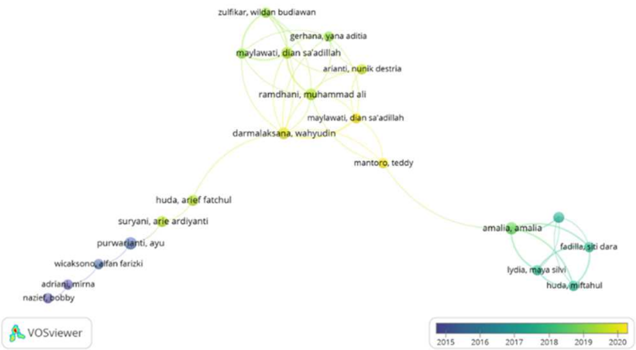


Figure 5: **Overlay Visualization**

In the lower right corner of Figure 5, there is a color bar indicating when a publication was made or published. The more to the right (the lighter) the newer, if the color is dark then the research is published in the previous time or longer. Dots/circles that have similar colors indicate the similarity of the year of publication.

Furthermore, the depth (density) of a document is shown in Figure 6, this depth is related to publication with many co-authors. The more co-authors, the brighter the density. If you only have one connection, it is blue.

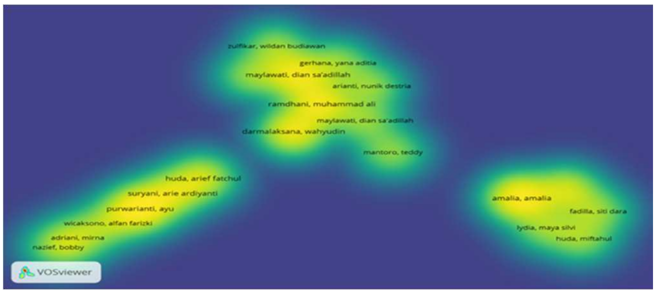


Figure 6: **Document Density**

2) *Co-Authorship with Unit Analysis: Countries*

In this analysis, the selected analysis unit is based on countries. The counting selected method is full counting. The threshold setting is set at number 5, in which the visualization shows countries which at least 5 documents in the existing data set. 3 countries have been successfully selected. They are Indonesia, Japan, and Malaysia as shown in Table 1.

Table 1. **Selected Countries**

Country	Documents	Citations	Total Link Strength
Indonesia	218	729	11
Malaysia	21	42	6
Japan	8	20	5

In the **Figure 7.** shows that the latest research or publications related to stemming are carried out more in Indonesia, then, Malaysia, and Japan. The bar color in the lower right corner indicates the year when the publication was made. It can be seen that Indonesian stemming publications have averaged publications in 2019, while Malaysia and Japan have gotten more in 2016 or 2017.

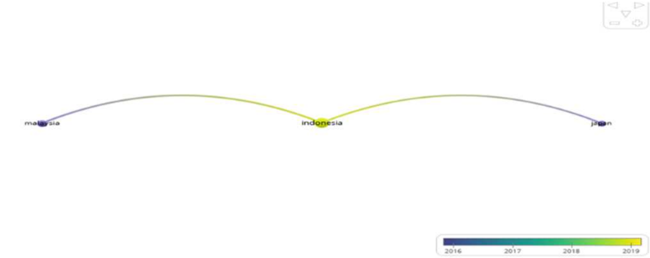


Figure 7: **Overlay Visualization for Countries Unit**

Density visualization in figure 8 shows that Indonesia's publications related to stemming are numerous, this is shown by a bulging oval in the yellow color. As shown in Table 1, authors from Malaysia and Japan published the publication of 21 documents and 8 documents.

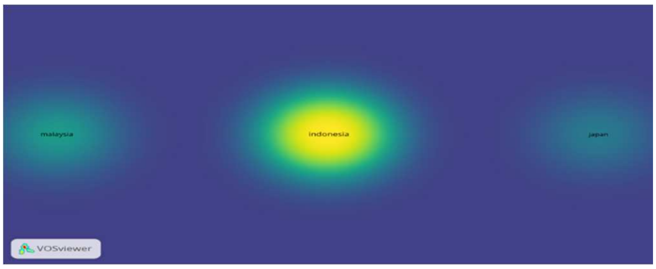


Figure 8 : **Density Visualisation for Countries Unit**

B. *Citation Engineering Analysis With Unit Analysis: Documents*

This citation technique will visualize the analyzed data (document) if a document cites other documents that are similar. Similar analyzed. The analysis unit is based on the unit document. Then, the minimum number of document citations is set to the number 1. The Visualization displays document citations in at least 1 document in the existing dataset. Once it is set up, of the 372 documents in the dataset, there are 196 matched and linked the data. Table 2 shows the author's document with the strongest link level. It needs to be noted that the number of citations obtained will not assure the level of link strength.

Table 2. **Selected Documents**

Documents	Citations	Total Link Strength
Adriani (2007)	98	56
Hidayat (2020)	1	5
Hasanah (2018)	8	5

Furthermore, of the 196 documents that were successfully selected, which had a link only 90 documents were shown in Figure 9. Where the visualization of the total strength of the document is indicated by a larger circle than the others. Different colors represent the total strength of each other study.

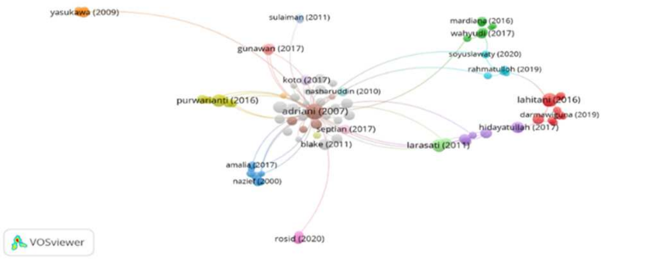


Figure 9: **Visualization of Document Citation**

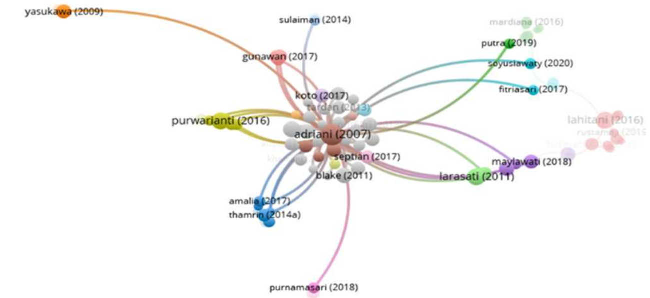


Figure 10: **Adriani's Document Citation**

Figure 10, highlighted in the Adriani document (2007), proved that the citation rate is the highest compared to the others (e.g. Lahtani (2016) and Purwarianti (2016) indicated by the number of available branches. It can be understood that Adriani's document (2007) became the basis for the development of the next stemming. Although Lahtani and Purwarianti have a citation lower rate than adriani, it can be seen in Figure 9 that both of them are derived from Adriani's documents, it is indicated that the branches connected both of them.

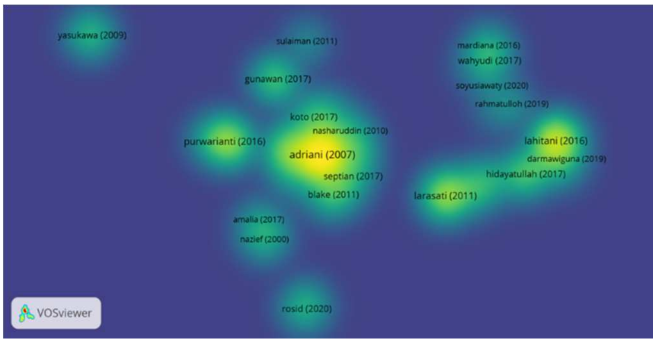


Figure 11: **Document Density**

The density document level is shown in figure 11. And further, it reinforces that Adriani's documents (2007) become a reference for development. Shown in yellow, it means that the citation rate is higher. from the picture, we can also see Lahtani (2016) has a fair large level of citations, followed by Purwarianti (2016).

IV. CONCLUSION

Stemming is an important step in the staging process while making the learning model machines, especially those involving Natural Language Processing. In particular, stemming has many algorithms which can be selected and applied following the existing problems. As time goes by and having complex problems, improvisations, and addition of new features have been added to existing algorithms. Nazief-Adriani's documents and their documents become a standard in the appearance of new algorithms such as Confix Stripping to Enhanced Confix Stripping.

Scientometric analysis shows that the stemming rate development is increasing from 2013 to 2022, and the development and implementation of the Nazief-Adriani algorithm have resulted in 310 publications. From the above analysis, it can be taken some conclusion that Nazief and ECS algorithm is still widely used for stemming in Indonesia language and continued to be developed. Figuring out from the citation analysis, Adriani's document (2007) explaining the Nazief-Adriani algorithm has been widely cited by other authors as the basis for the development of the continuality algorithm.

ACKNOLEDGEMENT

The research described in this paper was corporated with Laboratory of the Computer Science Binus University and Laboratory of The Computer Science Pakuan University

REFERENCES

[1] K. Baker, "XSTEM: An exemplar-based stemming algorithm," pp. 1–11, 2022, [Online]. Available: http://arxiv.org/abs/2205.04355.

[2] J. Singh and V. Gupta, *A systematic review of text stemming techniques*, vol. 48, no. 2. Springer Netherlands, 2017.

[3] M. S. H. Simarangkir, "Studi Perbandingan Algoritma - Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia," *J. Inkofar*, vol. 1, no. 1, pp. 40–46, 2017, doi: 10.46846/jurnalinkofar.v1i1.2.

[4] M. Haroon, "Comparative Analysis of Stemming Algorithms for Web Text Mining," *Int. J. Mod. Educ. Comput. Sci.*, vol. 10, no. 9, pp. 20–25, 2018, doi: 10.5815/ijmecs.2018.09.03.

[5] A. F. Hidayatullah, "The influence of stemming on Indonesian tweet sentiment analysis," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 2, no. August, pp. 127–132, 2015, doi: 10.11591/eecsi.v2i1.791.

[6] T. Winarti, J. Kerami, and S. Arief, "Determining Term on Text Document Clustering using Algorithm of Enhanced Confix Stripping Stemming," *Int. J. Comput. Appl.*, vol. 157, no. 9, pp. 8–13, 2017, doi: 10.5120/ijca2017912761.

[7] J. Asian, M. Adriani, B. Nazief, S. M. M. THAHAGHOGHI, and H. E. WILLIAMS, "Stemming Indonesian : A Confix-Stripping Approach," *Conf. Res. Pract. Inf. Technol. Ser.*, vol. 38, no. 4, p. 13, 2007, doi: 10.1145/1316457.1316459.

[8] A. Z. Arifin, P. A. D. . Mahendra, and H. T. Ciptaningtyas, "ENHANCED CONFIX STRIPPING STEMMER AND ANTS ALGORITHM FOR CLASSIFYING NEWS DOCUMENT IN Representation of Textual," *5thInternational Conf. Inf. Commun. Technol. Syst.*, pp. 149–158, 2009.

[9] R. Setiawan, A. Kurniawan, W. Budiharto, I. H. Kartowisastro, and H. Prabowo, "Flexible affix classification for stemming Indonesian Language," *2016 13th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2016*, 2016, doi: 10.1109/ECTICon.2016.7561257.

[10] I. Prismana, D. Prehanto, D. Dermawan, A. Herlingga, and S. Wibawa, "Nazief & Adriani Stemming Algorithm With Cosine Similarity Method For Integrated Telegram Chatbots With

Service," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1125, no. 1, p. 012039, 2021, doi: 10.1088/1757-899x/1125/1/012039.

[11] J. Jumadi, D. S. Maylawati, L. D. Pratiwi, and M. A. Ramdhani, "Comparison of Nazief-Adriani and Paice-Husk algorithm for Indonesian text stemming process," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1098, no. 3, p. 032044, 2021, doi: 10.1088/1757-899x/1098/3/032044.

[12] T. H. J. Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," *Procedia Comput. Sci.*, vol. 197, no. 2021, pp. 660–667, 2021, doi: 10.1016/j.procs.2021.12.187.

[13] D. Mustikasari, I. Widaningrum, R. Arifin, and W. H. E. Putri, "Comparison of Effectiveness of Stemming Algorithms in Indonesian Documents," *Proc. 2nd Borobudur Int. Symp. Sci. Technol. (BIS-STE 2020)*, vol. 203, pp. 154–158, 2021, doi: 10.2991/aer.k.210810.025.

[14] Y. D. Pramudita, S. S. Putro, and N. Makhmud, "Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 3, p. 269, 2018, doi: 10.25126/jtiik.201853810.

[15] I. O. Suzanti, A. Jauhari, N. Hidayanti, I. Y. Harianti, and F. A. Mufarroha, "Comparison of Stemming and Similarity Algorithms in Indonesian Translated Al-Qur ' an Text Search," *J. Ilm. KURSOR*, vol. 11, no. 2, pp. 91–100, 2021.

[16] I. Mulyana, "PENGEMBANGAN ALGORITMA STEMMING BAHASA INDONESIA MELALUI MODIFIKASI KELOMPOK IMBUHAN, URUTAN DAN CARA PENGHAPUSAN IMBUHAN BERBASIS MORFOFONEMIK," Gunadarma University, 2019.

[17] A. Mooghali, R. Alijani, N. Karami, and A. Khasseh, "Scientometric Analysis of the Scientometric Literature," *Int. J. Inf. Sci. Manag.*, vol. 9, no. 1, pp. 19–31, 2011.

[18] T. Tupan, R. N. Rahayu, R. Rachmawati, and E. S. R. Rahayu, "Analisis Bibliometrik Perkembangan Penelitian Bidang Ilmu Instrumentasi," *Baca J. Dokumentasi Dan Inf.*, vol. 39, no. 2, p. 135, 2018, doi: 10.14203/j.baca.v39i2.413.

[19] R. Rohanda and Y. Winoto, "Analisis Bibliometrika Tingkat Kolaborasi, Produktivitas Penulis, Serta Profil Artikel Jurnal Kajian Informasi & Perpustakaan Tahun 2014-2018," *Pustabiblia J. Libr. Inf. Sci.*, vol. 3, no. 1, p. 1, 2019, doi: 10.18326/pustabiblia.v3i1.1-16.

[20] B. de P. F. E Fonseca, R. B. Sampaio, M. V. de A. Fonseca, and F. Zicker, "Co-authorship network analysis in health research: Method and potential use," *Heal. Res. Policy Syst.*, vol. 14, no. 1, pp. 1–10, 2016, doi: 10.1186/s12961-016-0104-5.

[21] M. Sidiq, "Panduan Analisis Bibliometrik Sederhana," *J. Artic.*, no. June, 2019, doi: 10.13140/RG.2.2.15688.37125.