A Systematic Literature Review of Stemming in Non-Formal Indonesian Language

Yohana Karuniawati Paskahningrum*
Magister of Informatics Engineering
University of AMIKOM Yogyakarta
Jl. Ringroad Utara,
Yogyakarta, 55283, Indonesia

Ema Utami
Magister of Informatics Engineering
University of AMIKOM Yogyakarta
Jl. Ringroad Utara,
Yogyakarta, 55283, Indonesia

Ainul Yaqin
Computer Science Faculty
University of AMIKOM Yogyakarta
Jl. Ringroad Utara,
Yogyakarta, 55283, Indonesia

Abstract:- This study tries to review studies on the stemming process in non-standard Indonesian Language. This is done to understand the methods that researchers use to collect data from various sources, process the data that has been collected, and classify data so that it becomes information that is easier to understand. The researcher collected, filtered, and reviewed the discovered research papers using a Systematic Literature Review approach. We collect research works from ScienceDirect, IEEE, arXiv, ACM Digital Library, Semantic Scholar, Google Scholar, Springer, and Elsevier by choosing research published from 2016 to 2022. The purpose of the researcher conducting this literature review is to understand stemming in Indonesian, gain an understanding of data collection techniques, stemming methods, and study stemming results from previous research. This study collects and summarizes twenty-seven stemming studies on the Indonesian language, selected from the forty-seven previously collected studies. The study was conducted regarding how to collect data, the language-stemming research methods used, and the stemming research results.

Keywords:- *Stemming, Indonesian Language, Non-Standard, Natural Language Processing, Information Retrieval.*

I. INTRODUCTION

Indonesian as the national language of Indonesia is the most widely spoken language in the world. A survey said that in 2021, Indonesian was spoken by nearly 200 million people worldwide, so that it is in 10th place as the language with the most speakers[1]. Countries whose citizens can speak Indonesian include 239.000 people in Taiwan, 190.000 people in Hong Kong, 118.000 people in Singapore and 118.000 people in the Netherlands [2].

Many studies on the Indonesian language have been carried out [3] [4] [5] [6] [8]. One factor that interferes with the development of the Indonesian language is the influence of "slang" or "bahasa alay" which is a non-standard form in Indonesian. The use of slang in spoken language, SMS, Twitter, or in performances on stage and television, can still be tolerated. However, it turns out that the use of informal language in language activities such as writing and speaking is something that is often found in classrooms. The use of informal language is often found on test answer sheets, in student assignments and when presenting in front of the class, slang is still the champion in its use [5]. Social networks are media that are widely used by language speakers to communicate with each other over long distances via the internet. Social networks that are in great demand by the public are Instagram, Facebook, and Twitter [7].

Research that is more in-depth in the process of Indonesian stemming is [21] focused on a modified Idris stemmer (from Malay) to IN-Indris in the Indonesian context. From the experiment result, IN-Idris had an accuracy of approximately 82.81%. In a study conducted [24], data were collected from conversations between customer service representatives and consumers when booking airline tickets, namely OkeTiket via WhatsApp Messenger. The researcher found that Incorbiz chipped 85% and 73% more accurately than Sastrawi. However, research on Incorbiz is still ongoing and has not been completed by researchers. Unfortunately, this research has not been completed and is still ongoing today. A study [28] uses the Nazief-Adriani algorithm to detect similarities between 15167 slang and standard words. Results from a study of the Nazief-Adriani algorithm's use of text pre-processing and stemming show the most common distribution for slang and formal word similarity, with values ranging from 80% to 89.99%. A study [30] investigated how Indonesians often use “bahasa alay” when communicating on social media, such as Twitter. The informal affix-stemming algorithm is used by researchers in classifying words used by Indonesian people in social media. Words in Tweets are categorized as official, unofficial, and unknown. After doing the calculations, the researchers found that more than 12% of “bahasa alay” was used on social media.

The above studies have not yet been able to provide a complete overview of Indonesian language ancestry research, especially in relation to non-standard languages. I created a systematic literature review to answer it.

II. THE METHOD

A Systematic Literature Review (SLR) study aims to identify key relevant studies, extract the necessary data, and analyze and synthesize the results to gain broader insight into the research area [9].

Regardless of the field, discipline, or philosophical perspective, the author [10] stated that in order to perform SLR, it is necessary to perform six stages as below.

A. *Research Questions*

RQ1 : What data collection techniques do researchers use in stemming research?
RQ2 : What are the methods used in stemming research?
RQ3 : What are the stemming results in the research?

B. *Research Strategy*

The researcher looking for papers from ScienceDirect, IEEE, Springer, arXiv, ACM Digital Library, Semantic Scholar, Google Scholar, and Elsevier using two keywords, including words in Indonesian and also in English:

- “stemming” and “non-standard Indonesian language”
- “bahasa Indonesia tidak baku” atau “stemming bahasa Indonesia”

C. *Study Selection*

Criteria should be established when evaluating manuscripts. The researcher has two types of criteria that can be used in writing papers: inclusion criteria and exclusion criteria. Here are some inclusion criteria for this study:

- The research collected is research conducted from 2016 to 2022.
- The researcher selects papers that are written in Indonesian and English.
- The study main topic must be Indonesian stemming

The exclusion criteria for this study were:

- Research that is not included in the inclusion criteria.
- This research is not clear in describing the research flow and how to conduct research
- Research that fails to address the research objectives.

Fig 1 shows the criteria in this study.

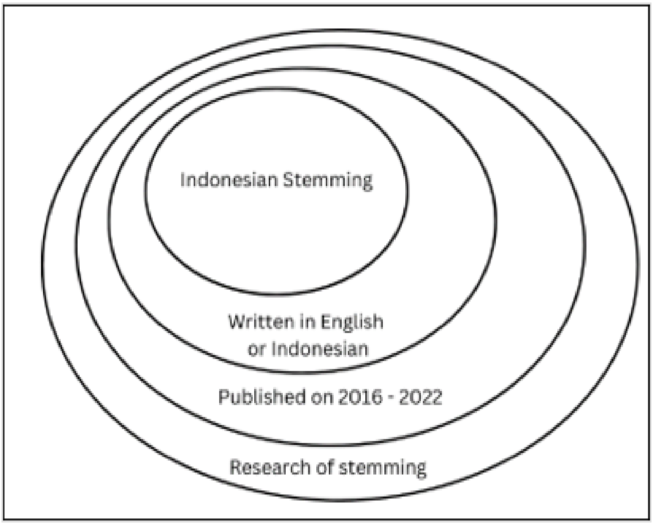

Fig. 1 Criteria in this research

The data selection process is depicted in Fig 2.

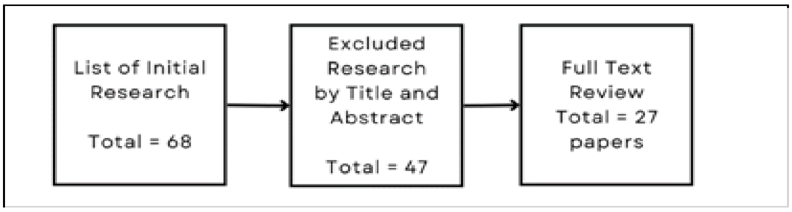
List of Initial Research
Total = 68
Excluded Research by Title and Abstract
Total = 47
Full Text Review Total = 27 papers

Fig. 2The data selection

D. *Quality Assessment*

A quality assessment should be carried out to get a clear picture of the quality of a study. Quality assessment is carried out to make decisions regarding whether the data was found to be used in this study or not [17]. In this study, the data that has been found will then be evaluated using the question of the quality assessment criteria below:

- Was the research published from 2016 to 2022?
- Whether the research is written in Indonesian or English?
- Whether the research topic is Indonesian stemming?

If all the questions above are met, then the research can be continued in the literature review process.

E. *Data Extraction*

At this point, the data extracted from the reviewed paper is in what year the research was published, the data used in the reviewed research, and data collection techniques in the reviewed research, the method used in the stemming process in the reviewed research, and the results of stemming on the research. reviewed research. Then the researcher entered all the data into a spreadsheet document and the researcher would synthesize the data [18].

F. *Data Synthesizing*

At this stage, a total of 68 studies have been collected and selected based on the title and abstract. From this process, 47 papers were produced. Furthermore, these 47 papers were selected using inclusion and exclusion criteria. If it meets the inclusion criteria, it can be used in this literature review. If the study meets the exclusion criteria, it will not be used in this literature review. From this process, 27 papers were obtained which were finally reviewed and analyzed. The data taken from papers and main findings are analyzed and consolidated in Table 1

III. RESULT AND DISCUSSION

A. *The Data Collection Techniques*

In conducting a study, researchers need data as research objects. The process of collecting and analyzing accurate data from various sources to find answers to research questions, trends, possibilities, etc., and to evaluate possible outcomes is called data collection. During data collection, researchers should identify the type of data, data sources, and methods used. It quickly becomes apparent that there are many different ways to collect data. The research, commercial, and government sectors rely heavily on data collection.

Previous researchers used a variety of different data collection techniques. Research [21], [33], [35], [36], [42], [44] collect the data from printed documents, such as novels or articles.

Research [23], [24], [25], [26], [27], [28], [30], [31], [32], [39], [41], [43], [45], [46], [47] collect data from web, like Twitter, detik.com, Republika.co.id and so on. Data words from a corpus or dictionary were used in research [22], [29] [34], [37], [38], [40]. An overview of data collection techniques can be understood by looking at Fig 3.

Table 1 Reviewed Paper (A)

Research	Year	Language	Data collection
[21]	2022	Indonesian	printed documents
[22]	2022	Indonesian	corpus or dictionary
[23]	2021	Indonesian	data from web
[24]	2021	Indonesian	data from web
[25]	2021	Indonesian	data from web
[26]	2021	Indonesian	data from web
[27]	2020	Indonesian	data from web
[28]	2020	Indonesian	data from web
[29]	2020	Indonesian	corpus or dictionary
[30]	2019	Indonesian	data from web
[31]	2019	Indonesian	data from web
[32]	2019	Indonesian	data from web
[33]	2019	Indonesian	printed documents
[34]	2018	Indonesian	corpus or dictionary
[35]	2020	Nepali	printed documents
[36]	2018	Indonesian	printed documents
[37]	2018	Indonesian	corpus or dictionary
[38]	2018	Indonesian	corpus or dictionary
[39]	2018	Indonesian	data from web
[40]	2018	Indonesian	corpus or dictionary
[41]	2017	Indonesian	data from web
[42]	2017	Indonesian	printed documents
[43]	2017	Indonesian	data from web
[44]	2016	Indonesian	printed documents
[45]	2016	Indonesian	data from web
[46]	2016	Indonesian	data from web
[47]	2016	English	data from web

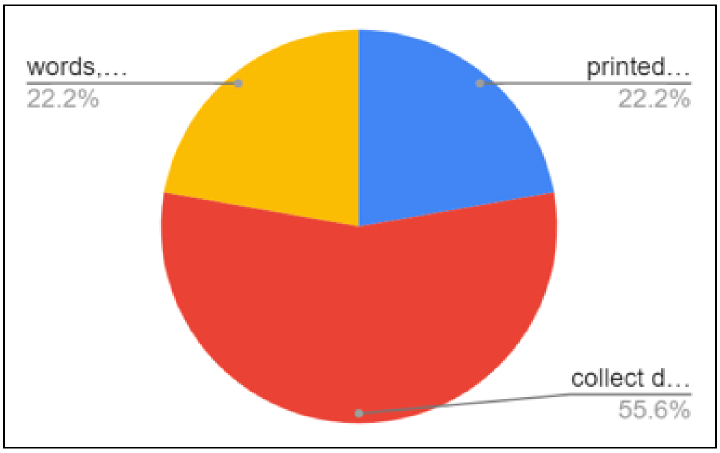

Fig. 3 The data collection techniques

Table 2 Reviewed Paper (B)

Research	Stemming Method	Stemming Result
[21]	In-Idris Algorithm	82.81
[22]	CSFNNG2P	84.37
[23]	Nazief-Adriani	86.91
[24]	Sastrawi	85
[25]	Sastrawi	85
[26]	CBOW	91
[27]	Nazief-Adriani	92.4
[28]	Nazief-Adriani	90
[29]	BoW	67.3
[30]	Non-formal Affixed Stemmer	not in number
[31]	Enhanced Confix Stripping; Sastrawi	98.45
[32]	Proposed Stemming	64.8
[33]	Enhanced Confix Stripping	93
[34]	Jaro-Winkler	85
[35]	Rule Based Stemmer	not in number
[36]	Spellchecker	98
[37]	Flexible Affixed Classification	73.3
[38]	Non-formal Affixed Stemmer; Levenshtein Distance	96.6
[39]	Nazief-Adriani; Porter	88.65
[40]	Cosine Similarity	not in number
[41]	Nazief-Adriani; Confix Stripping	79.45
[42]	Incremental Stemming	79.12
[43]	Nazief-Adriani	99.76
[44]	Nazief-Adriani	96.46
[45]	Flexible Affixed Classification	97.38
[46]	Nazief-Adriani	90.91
[47]	Classical English stemmer	not in number

B. *The Stemming Method*

Stemming is the process of reducing inflections or derivations to their root forms (words or roots), much like the derivation "comfortable" is reduced to its root "comfort". This does not necessarily mean reducing the word to its dictionary stem. It uses some algorithm to decide how to truncate words. This is the main difference between word stemming and lemmatization which reduces words to dictionary roots which are more complex and require a very high level of linguistic competence.

Stemming in Indonesian is quite difficult, because Indonesian is known to have 127,000 basic words recorded in the Big Indonesian Dictionary. Stemming is the process of finding root words from affixed words by removing all affixes consisting of prefixes, infixes, suffixes and combinations of prefixes and suffixes. .stemming is one of the important processes that greatly affect the quality of the analysis results. There are many algorithms used to carry out the stemming process, including the Nazief-Andriani algorithms.

Previous research has used different algorithms for stemming [11][19][20]. [21] used the In-Idris Algorithm in processing 10 Indonesian documents. A method called CSFNNG2P (Combined Confix Stripping Based Stemming And Fuzzy Nearest Neighbor Based Graphene To Phoneme) conversion was used by research [22] to process the 50.000 Indonesian words and converted into more than 300.000 unique words. The most widely used algorithm for stemming research is the Nazief-Adriani algorithm used in research [23], [28], [39], [41], [43], [44], [46]. While Sastrawi is used in several studies such as [24], [25], [31].

Sastrawi [17] is a simple module owned by the python library that allows the reduction of inflected words in Indonesian to their standard form or according to dictionary standards [12] [13]. Research [30] [38] uses the Non-formal Affixed Stemmer algorithm and uses the Enhanced Confix Stripping algorithm (ECS) [31] [33]. The Proposed Stemming Algorithm was used by research [32] to process data from Twitter as many as 2.7345 tweets. Meanwhile, research [34] uses the Jaro-Winkler algorithm in stemming with the help of the Levenshtein Distance algorithm. Spellchecker stemming research was used [36] to perform stemming on 5 novels and 5 articles.

The Flexible Affixed Classification algorithm was used in the study [37] [45]. Research [38] applies the Levenshtein Distance algorithm to process 60 non-standard Indonesian words. Meanwhile, Porter's algorithm was used in research [39] to perform stemming on 379 data, while also using the Lucene algorithm in the same study. In contrast to ECS, the Confix Stripping algorithm was used in the study [41] with 2,141,421 words, from those that were got 79.453 unique words. The Incremental Stemming algorithm was used in research [42] in optimizing stemming on 6,464 text data. The algorithm that has been used for a long time is the Classical English stemmer in research [47] to examine datasets taken from major sites abroad. The stemming techniques are shown in Fig 4.

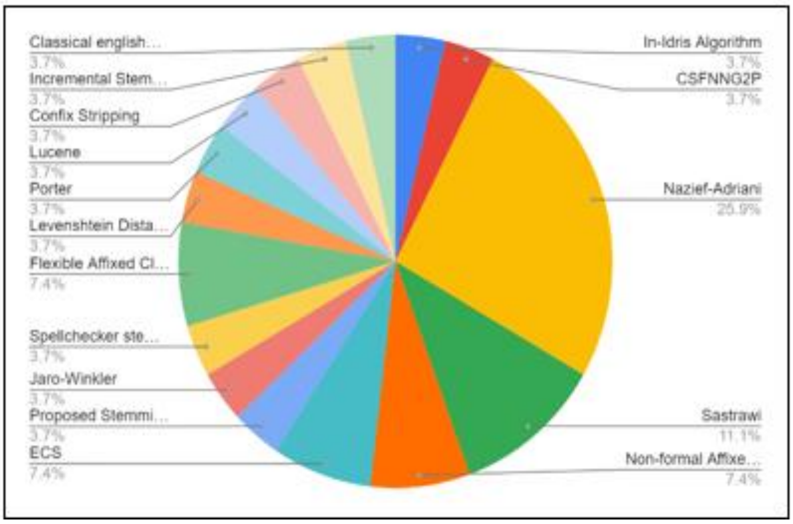

Fig. 4 The stemming method

C. *The Stemming Result*

In the application of the stemming process to Indonesian documents, there are three types of errors, namely understemming, which means that the word has been stemmed too little, overstemming, which means that the word has been stemmed too much, and unchanged, which means that the word has not changed at all. Understemming is the lack of beheading, overstemming is beheading, whereas unchanged reminds the same word without change[14][15][16]. The cause of failure in the test is because there are spelling exception cases and overstemming cases. Spelling exception is a case where the root word still has an affix after passing through the stemming process. Meanwhile, overstemming is a case where the basic word has decreased letters or syllables which are considered as affixes as a result of the stemming process [17].

The results of stemming research are in the 71%-75% range, namely [24],[25],[37]. In detail, studies [24] and [25] both obtained 73% results. Research [37] obtained a yield of 73.3%. Research results that are in the range of 75%-80% are studies [41],[42]. Research [41] obtained a yield of 79.45% and research [42] obtained 79.12%. Research [21],[22],[24],[25] obtained stemming results in the range of 81%-85%. Research [21] obtained a yield of 82.81%, research [22] obtained a yield of 84.37%. While research [24] and [25] both obtained 85% results. The results of the study with a range of 86%-90%, namely [23],[28],[39]. Research [23] obtained a yield of 86.91%. Research [28] obtained 90% of the research he conducted, while [39] obtained results of 88.65%. For research with range of 91%-95%, namely [27],[33],[46]. In detail, research [27] obtained a yield of 92.4%. Research [33] obtained 93% results, and research [46] obtained 90.91% in their researches. The best results from the studies reviewed were studies [36],[38],[43],[44],[45] with a range of results close to 100%. It is stated in detail that research [36] obtained a result of 98%, research [38] obtained a result of 96.6%. Unlike the two, research [43] obtained near-perfect results, namely 99.76, while research [44] obtained results of 96.46 and research [45] obtained results of 97.38%. The results of stemming research are expressed as a percentage of the number of words that have been successfully stemming.

Table 3 Overview of the Stemming Result

Research	Stemming Result
[29],[32]	under 70%
[21],[24],[25],[37]	71-75%
[41],[42]	75%-80%
[21],[22],[24],[25]	81%-85%
[23],[28],[39]	86%-90%
[27],[33],[46]	91%-95%
[36],[38],[43],[44],[45]	96%-100%
[30],[35],[40],[47]	unknown

Research [30] [35] [40] did not mention the percentage because the study did not measure the accuracy of the stemming process and the study [47] was a literature review study. The stemming results are shown in Fig 5 .

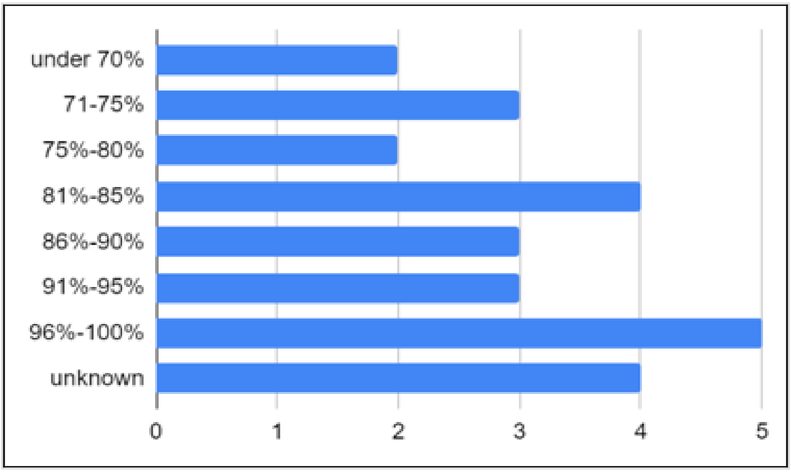

Fig. 5 The stemming result

IV. CONCLUSION

This study brings together various stemming studies on Indonesian and other languages. After reviewing the titles, abstracts, and content of studies found in ScienceDirect, IEEE, Springer, arXiv, ACM Digital Library, Semantic Scholar, Google Scholar, and Elsevier, 27 studies were selected for further consideration. A review was conducted on data collection techniques, stemming methods, and the results of stemming studies. The conclusion of this review is that the most widely used method for stemming the Indonesian language is Nazief-Adriani. This is because the method that gets the best results for non-standard Indonesian stemming is by using the Nazief-Adriani stemming method. In addition, the most frequently used data source is retrieving data from websites such as social media.

Because this research does not examine the stemming method in depth, further research is needed, particularly a review of the literature on "Alay language" stemming from a wider variety of sources. It is also advisable to do stemming research on the use of Alay language from private chat applications such as WhatsApp, both published in chats and on statuses. This study only obtained 27 research reports to be studied as literature review work. Further research would be better to use more research reports to be studied into a more detailed and in-depth literature review.

REFERENCES

[1]. Kasih, A.P., “10 Bahasa Paling Banyak Digunakan di Dunia, Indonesia Nomor Berapa?” *KOMPAS.com*, 2021. [Online]. Available: https://www.kompas.com/edu/read/2021/08/05/162355371/10-bahasa-paling-banyak-digunakan-di-dunia-indonesia-nomor-berapa?page=all [Accessed: 17-Oct-2022]

[2]. Devlin, T. M., "How Many People Speak Indonesian, And Where Is It Spoken?, Babbel Magazine, 2021. [Online]. Available: https://www.babbel.com/en/magazine/how-many-people-speak-indonesian-where-is-it-spoken [Accessed: 17-Oct-2022]

[3]. Puspitasari, Andi. "Menumbuhkan bahasa Indonesia yang baik dan benar dalam pendidikan dan pengajaran." Tamaddun 16, no. 2 (2017): 81-87.

[4]. Muchti, Andina, And Yeni Ernawati. "Penguasaan Kosakata Baku Dan Tidak Baku: Sebuah Studi Kasus Mahasiswa Ubd." Jurnal Ilmiah Bina Edukasi 15, no. 1 (2022): 61-70.

[5]. Purwarianti, A., Andhika, A., Wicaksono, A. F., Afif, I., & Ferdian, F. (2016). InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification. 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA). doi:10.1109/icaicta.2016.7803103

[6]. Lakonawa, Katarina N., Sebastianus AS Mola, and Adriana Fanggidae. "Nazief-Adriani Stemmer Dengan Imbuhan Tak Baku Pada Normalisasi Bahasa Percakapan Di Media Sosial." Jurnal Komputer dan Informatika 9, no. 1 (2021): 65-73..

[7]. Gustiasari, Dewi Rani. "Pengaruh perkembangan zaman terhadap pergeseran tata Bahasa Indonesia; Studi kasus pada pengguna instagram tahun 2018." Jurnal Renaissance 3, no. 2 (2018): 433-442.

[8]. Abidin, Jenal. "Pembangunan Kamus Bahasa Indonesia Kata Tidak Baku Menggunakan Algoritma Letent Semantic Indexing Dan Damerau Levenshtein Distance." PhD diss., Universitas Komputer Indonesia, 2017.

[9]. van Dinter, Raymon, Bedir Tekinerdogan, and Cagatay Catal. "Automation of systematic literature reviews: A systematic literature review." Information and Software Technology 136 (2021): 106589.

[10]. Durach, Christian F., Joakim Kembro, and Andreas Wieland. "A new paradigm for systematic literature reviews in supply chain management." Journal of Supply Chain Management 53, no. 4 (2017): 67-85.

[11]. Suyanto, Suyanto, Andi Sunyoto, Rezza Nafi Ismail, Ema Rachmawati, and Warih Maharani. "Stemmer and phonotactic rules to improve n-gram tagger-based Indonesian phonemicization." *Journal of King Saud University-Computer and Information Sciences* 34, no. 6 (2022): 3807-3814.

[12]. Siswanto, Boby, and Yasi Dani. "Sentiment Analysis about Oximeter as Covid-19 Detection Tools on Twitter Using Sastrawi Library." In *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, pp. 161-164. IEEE, 2021.

[13]. Widodo, " "NLP Sederhana Dengan Python", Fani Widodo, 2021. [Online]. Available: https://sites.unpad.ac.id/widodo/2021/03/09/nlp-dengan-python/ [Accessed: 17-Oct-2022]

[14]. Srinidhi, Sunny. "Stemming of words in Natural Language Processing, what is it?",Towards Data Science, 2020. [Online]. Available: https://towardsdatascience.com/stemming-of-words-in-natural-language-processing-what-is-it-41a33e8996e2 [Accessed: 17-Oct-2022]

[15]. Jain, Saurav, "Introduction to Stemming", Geeks for Geeks, 2021. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-stemming/ [Accessed: 17-Oct-2022]

[16]. Putra, Syopiansyah, Novi Cahyanti, Suci Ratnawati, Muhamad Gunawan, and Dwi Sari. "The Implementation of Indonesian Stemming System for Indonesian Translation of the Quran." In *Proceedings of the 2nd International Conference on Quran and Hadith Studies Information Technology and Media in Conjunction with the 1st International Conference on Islam, Science and Technology, ICONQUHAS & ICONIST, Bandung, October 2-4, 2018, Indonesia*. 2020.

[17]. Guterres, Anita, Magister Teknologi Infomasi, and Joan Santoso. "Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based." *Teknika* 8, no. 2: 142-147.

[18]. Pradana, Aditya Wiha and Mardhiya Hayaty. "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts." *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control* (2019)

[19]. Simarangkir, Manase Sahat H. "Studi Perbandingan Algoritma-Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia." *Jurnal Inkofar* 1, no. 1 (2017).

[20]. R. N. Ismail and S. Suyanto, "Indonesian Graphemic Syllabification Using n-Gram Tagger with State-Elimination," 2020 8th International Conference on Information and Communication Technology (ICoICT), 2020, pp. 1-6, doi: 10.1109/ICoICT49345.2020.9166368

[21]. Suci, Febiarty Wulan, Nur Hayatin, And Yuda Munarko. "In-Idris: Modification Of Idris Stemming Algorithm For Indonesian Text." *Iium Engineering Journal* 23, No. 1 (2022): 82-94.

[22]. Dwiharyono, Hari, and Suyanto Suyanto. "Stemming for Better Indonesian Text-to-Phoneme." *Ampersand* 9 (2022): 100083.

[23]. Hidayat, Wahyu. Analisis Penerapan Algoritma Stemming Nazief & Adriani Pada Perbandingan Algoritma Winnowing Dan Algoritma Ratcliff/Obershelp Pada Pendeteksi Kesamaan Produk Di Marketplace. Diss. Universitas AMIKOM Yogyakarta, 2021.

[24]. Mutiara, Achmad Benny, Eri Prasetyo Wibowo, and Paulus Insap Santosa. "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation." *Journal of Big Data* 8, no. 1 (2021): 1-16.

[25]. Santosa, Paulus Insap. "Improving Stemming Techniques For Non-Formal Indonesian Sentences Using Incorbiz."

[26]. Riyaddulloh, Riri, and Ade Romadhony. "Normalisasi Teks Bahasa Indonesia Berbasis Kamus Slang Studi Kasus: Tweet Produk Gadget Pada Twitter." *eProceedings of Engineering* 8, no. 4 (2021).

[27]. Mustikasari, Dyah, Ida Widaningrum, Rizal Arifin, and Wahyu Henggal Eka Putri. "Comparison of Effectiveness of Stemming Algorithms in Indonesian Documents." In *2nd Borobudur International Symposium on Science and Technology (BIS-STE 2020)*, pp. 154-158. Atlantis Press, 2021.

[28]. Hidayat, Wahyu, Ema Utami, and Anggit Dwi Hartanto. "Effect of Stemming Nazief & Adriani on the Ratcliff/Obershelp algorithm in identifying level of similarity between slang and formal words." In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 22-27. IEEE, 2020.

[29]. Utomo, Fandy Setyo, Nanna Suryana, and Mohd Sanusi Azmi. "Stemming Impact Analysis On Indonesian Quran Translation And Their Tafsir Classification For Ontology Instances." IIUM Engineering Journal 21, no. 1 (2020): 33-50.

[30]. Utami, Ema, Anggit Dwi Hartanto, Sumarni Adi, Rahardyan Bisma Setya Putra, and Suwanto Raharjo. "Formal and non-formal Indonesian word usage frequency in twitter profile using non-formal affix rule." In 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS), vol. 1, pp. 173-176. IEEE, 2019.

[31]. Yusliani, Novi, Rifkie Primartha, and Mastura Diana Marieska. "Multiprocessing Stemming: A Case Study of Indonesian Stemmi." International Journal Computer and Applications (IJCA) 182, no. 40 (2019): 15-19.

[32]. Rizki, Afian Syafaadi, Aris Tjahyanto, and Rahmat Trialih. "Comparison of stemming algorithms on Indonesian text processing." TELKOMNIKA (Telecommunication Computing Electronics and Control) 17, no. 1 (2019): 95-102.

[33]. Rifai, Wafda, and Edi Winarko. "Modification of Stemming algorithm using a non-deterministic approach to Indonesian text." IJCCS (Indonesian Journal of Computing and Cybernetics Systems) 13, no. 4 (2019): 379-388.

[34]. Qulub, Mudawil, Ema Utami, and Andi Sunyoto. "Stemming Kata Berimbuhan Tidak Baku Bahasa Indonesia Menggunakan Algoritma Jaro-Winkler Distance." Creative Information Technology Journal 5, no. 4 (2020): 254-263.

[35]. Koirala, Pravesh, and Aman Shakya. "A Nepali Rule Based Stemmer and its performance on different NLP applications." *arXiv preprint arXiv:2002.09901* (2020).

[36]. Syawanodya, Indira, and Arief Fatchul Huda. "Improvement on Stemmer Algorithm for Indonesian Language With Spellchecker." In 2018 Third International Conference on Informatics and Computing (ICIC), pp. 1-5. IEEE, 2018.

[37]. Putra, Rahardyan Bisma Setya, and Ema Utami. "Non-formal affixed word stemming in Indonesian language." In 2018 International Conference on Information and Communications Technology (ICOIACT), pp. 531-536. IEEE, 2018.

[38]. Putra, Rahardyan Bisma Setya, Ema Utami, and Suwanto Raharjo. "Optimalisasi Stemming Kata Berimbuhan Tidak Baku Pada Bahasa Indonesia Dengan Levenshtein Distance." Jurnal Informatika: Jurnal Pengembangan IT 3, no. 2 (2018): 200-205.

[39]. Maylawati, Dian Sa'adillah, Wildan Budiawan Zulfikar, Cepy Slamet, Muhammad Ali Ramdhani, and Yana Aditia Gerhana. "An improved of stemming algorithm for mining indonesian text with slang on social media." In 2018 6th International Conference on Cyber and IT Service Management (CITSM), pp. 1-6. IEEE, 2018.

[40]. Hasanah, Uswatun, Tri Astuti, Rizki Wahyudi, Zanuar Rifai, and Rilas Agung Pambudi. "An experimental study of text preprocessing techniques for automatic short answer grading in Indonesian." In 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE), pp. 230-234. IEEE, 2018.

[41]. Widayanto, Hari, and Arief Fatchul Huda. "Comparison Nazief Adriani and CS stemmer algorithm for stemm real data." e-Proceeding of Engineering 4, no. 3 (2017): 5215.

[42]. Hidayat, Wahyu. "Ekstraksi Kata Dasar Secara Berjenjang (Incremental Stemming) Berbasis Aturan Morfologi untuk Teks Berbahasa Indonesia." Jurnal Infotel 9, no. 2 (2017): 166-171.

[43]. Prihatini, Putu Manik, IKG Darma Putra, I. A. D. Giriantari, and M. Sudarma. "Stemming Algorithm for Indonesian Digital News Text Processing." International Journal of Engineering and Emerging Technology 2, no. 2 (2018): 1-7.

[44]. Mardiana, Tari, Teguh Bharata Adji, and Indriana Hidayah. "Stemming influence on similarity detection of abstract written in Indonesia." TELKOMNIKA (Telecommunication Computing Electronics and Control) 14, no. 1 (2016): 219-227.

[45]. Setiawan, Reina, Aditya Kurniawan, Widodo Budiharto, Iman Herwidiana Kartowisastro, and Harjanto Prabowo. "Flexible affix classification for stemming Indonesian Language." In 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 1-6. IEEE, 2016.

[46]. Hidayatullah, Ahmad Fathan, Chanifah Indah Ratnasari, and Satrio Wisnugroho. "Analysis of stemming influence on indonesian tweet classification." TELKOMNIKA (Telecommunication Computing Electronics and Control) 14, no. 2 (2016): 665-673.

[47]. Singh, Jasmeet, and Vishal Gupta. "A systematic review of text stemming techniques." Artificial Intelligence Review 48, no. 2 (2017): 157-217.