JOIN (Jurnal Online Informatika)
Volume 7 No. 2 | December 2022: 184-192
DOI: 10.15575/join.v7i2.846
p-ISSN: 2528-1682
e-ISSN: 2527-9165

Multi Rule-based and Corpus-based for Sundanese Stemmer

Ade Sutedi[1], Muhammad Rikza Nasrulloh[2], Rickard Elsen[3]
[1,2,3]Department of Informatics Engineering, Institut Teknologi Garut, Indonesia

Article Info

Article history:

Received December 22, 2021
Revised June 15, 2022
Accepted June 21, 2022
Published December 26, 2022

Keywords:

Corpus-based
Multi Rule-based
Stemmer
Sundanese

ABSTRACT

The purpose of this study is to develop a stemming method by involved several methods including morphological (with affix and pro-lexeme removal), syllable (canonical) pattern, and corpus data as a comparison of the final results of stemming. The algorithm checks a number of the string first and removes affixes, then check the syllable pattern according to the stripping result, then compares to the corpus data which determines the final stemming process. In this study, the corpus data was taken from Sundanese dictionary consists of a single word used for the root word and the extracted dataset from the online Sundanese magazine. The results showed that the stripping of affix and pro-lexeme can remove the corresponding affixes and pro-lexeme then compares words that have a syllable pattern then executes the basic words quickly and the use of corpus can improve accuracy and reduce the over-stemming problems that occur in the stemming process.

Corresponding Author:

Ade Sutedi
Department of Informatics Engineering,
Institut Teknologi Garut,
Jl. Mayor Syamsu No. 1 Jayaraga, Tarogong Kidul, Garut, Indonesia
Email: adesutedi@itg.ac.id

1. INTRODUCTION

Big data becomes a new valuable resource since a human interact through social media and other platforms. There are consists of various kinds of unstructured information resources, ranging from text, sound, images, links, document files, videos, and etc. In order to utilize these resources, we need to find the necessary information that becomes more effective. This condition motivates the emergence of Artificial Intelligence (AI), Data Mining (DM), Machine Learning (ML), Natural Language Processing (NLP), and any other technologies to take benefits from the data spread out both online and offline. All of these applications generally use stemming techniques. Stemming is the process of returning derivation of word into basic or root word by removing particle of affixes (prefix, infix, suffix, and confix) formed that generally used in the searching task, Information Retrieval (IR), and etc.

The most popular stemming method that widely adopted and become the standard approach for Information Retrieval (IR) is Porter's method [1]. Furthermore, stemming application for the Indonesian language was developed in [2-6] has been successful. Generally, the state of the art from the previous research involves the use of the rules-based for affixes removal algorithms, corpus-based feature, and statistical techniques. The stemming technique in [2] focused on confix stripping approach for derived words to their morphological roots. In order to handle the ambiguity of word root result, the research in the [3] proposed the nondeterministic method with rule of disallowed prefix-suffix combination and word list references to give an alternative result after the stemming process. The research in the [4] claimed that the Paice/ Husk iterative algorithm could improve the Porter's stemmer algorithm that have over-stemming and under-stemming problem. The algorithm used in [2] and another Indonesian stemmer was compare and modified some rule from the previous Indonesian stemmer algorithm [5]. Another improvement algorithm for stemming Indonesian is using infix stripping [6]. This idea arose because in previous studies it did not include a rule for infix stripping [6] that widely used in Indonesian words. All of these methods have successfully applied for the stemming process with good results. Meanwhile, stemming development for regional languages is still unpopular because of the limited data resources and the habit of people who used national languages prioritized.

This research is driven by the lack of corpus data availability and the development activities in software tools used for various applications, especially those related to the Sundanese language.

The recent studies that discuss stemming techniques for regional languages have begun such as Javanese [7,8], Madura [9], Balinese [10,11], and Sundanese [12-15]. The stemming technique in previous research was developed in morphology form including affix derivation and inflection removal for English [1], Indonesian [2-6], and regional language [7-15]. The use of confix stripping is one of the techniques used in Javanese language stemming [7]. This technique adapted from the [2] algorithm and then developed into Enhanced Confix Striping (ECS). Likewise with the study [8] where the stemming method used was Removal by making the rule according to the Javanese language. In line with the study [7], ECS was also developed for Madura language stemmer [9]. Here, the confix stripping modification is done by changing the rule that adjusted to the form of Madura's morphology. Furthermore, the development of Balinese language is used by Porter's Stemmer [10] and a combination of Rule-Based Affix Removal with Statistic N-Grams [11] which is applied in specific cases.

However, the problems that occur in almost of previous research are under-stemming [2-5,7], over-stemming [2,6,7,13,15], affixes removal problem [3,8], and corpus limitation [7,10,12,13,15]. So, we try to improve the stemming problem with the adaptation of previous research and combine with the state of the art that we proposed. This research aims to use a multi rule-based and corpus-based method in handling the stemming process.

2. METHOD

In linguistics, there are rules used to determine the method of forming a word or a sentence. These methods such as morphology, phonology, syntactical, etc. The use of morphological concept is also used in other stemming research such as for English language [1], Indonesian language [2-4], and regional languages [5,6-9,12-15]. Furthermore, the use of corpus data is also widely applied in similar studies as comparative data if the process of stemming cannot be resolved, for example in research [2-4,6-11]. Another technique was developed with statistic for stemming with N-gram feature [11]. The Figure 1 describes the flow of the stemming concept that applied to several previous studies using the concept of rule, corpus, and statistics.

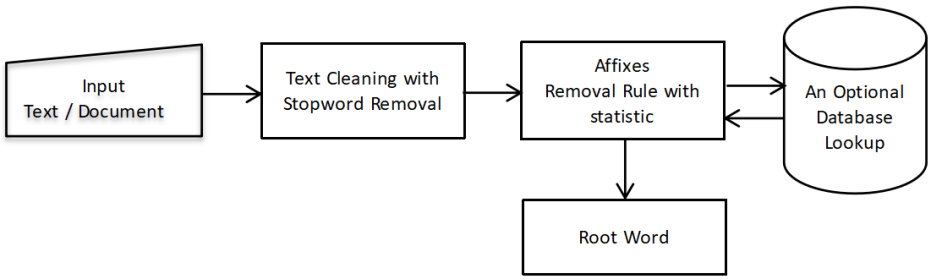

Figure 1. General Method for Stemming

In this study, the stemming process involved several rules such as morphology (affixes removal for derived word), phonology with syllable pattern, and the use of corpus data as a root word dictionary. This concept adopted because the Sundanese language has a grammar including sound, words, and characters [20]. So, the process aims to solve the problems that found in the stemming process especially in Sundanese stemmer [12-14]. The phonology will use to determine the canonical or syllable pattern in a word that is composed of vowels and consonants and count the number of characters in the word. The morphology in this study focuses on how to cut affixes in a words. The corpus data is used to compare the results of the affix cutting process and compare the corresponding syllable patterns. The improved of stemming process in this research depicted on Figure 2.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

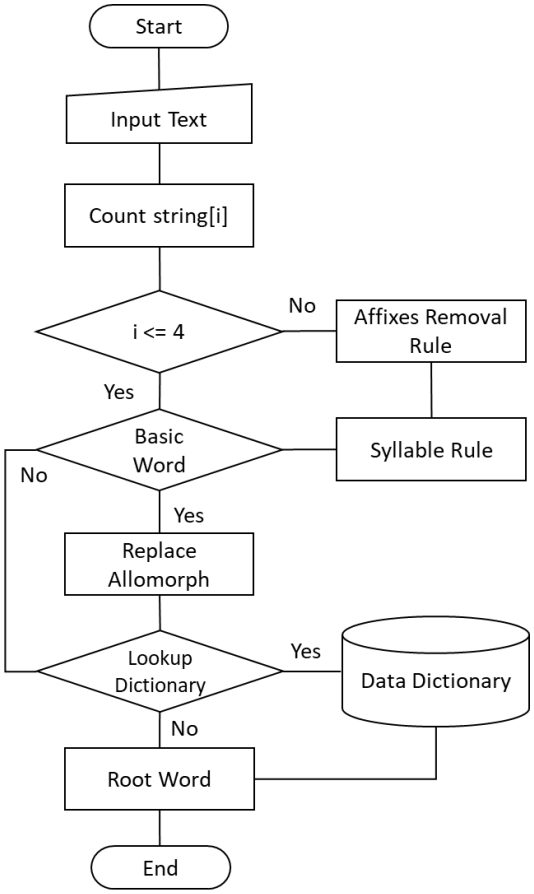

Figure 2. The Improved Method for Sundanese Stemming

2.1. Phonology

Sundanese language recognizes 25 phonemes which are divided into two types of segmental phonemes including 7 vowel phonemes and 18 consonant phonemes [16]. However, along with the development of writing and grammar in Sundanese language there were 5 consonant phonemes which were a variation of the original consonant phoneme [17]. From these phonemes, the Sundanese word has a syllable form and a canonical form which has the following rules:

1. The syllable form are as follows:

a.	***Ekasuku***	*a* (V)	big brother
		ka (KV)	to
		tuh (KVK)	there
b.	***Dwisuku***	*ua* (V-V)	uncle
		adi (V-KV)	little brother
		istri (VK-KKV)	wife, woman
c.	***Trisuku***	*iasa* (V-V-KV)	can
		awewe (V-KV-KV)	girl
		kalapa (KV-KV-KV)	coconut
d.	***Catursuku***	*atanapi* (V-KV-KV-KV)	or
		jatukrami (KV-KV-KKV-KV)	married
		jangjawokan (KVK-KV-KV-KVK)	spell, incantation
e.	***Pancasuku***	*balakasikang* (KV-KV-KV-KV-KVK)	boyish woman
		nengterewelang (KVK-KV-KV-KV-KVK)	talk fast and fluent

2. a consonant cannot stand alone as a syllable;
3. vowels can stand alone as syllables;
4. consonant clusters only occur at the beginning of the syllable;
5. The second consonant in a consonant group generally consists of consonants **l** , **r**, and **y** are very few in number
6. The arrangement of the phoneme *ekasuku* is a basic pattern of words with two or more syllables

Based on the syllable pattern in the Sundanese language, the basic word formed from the *ekasuku* pattern consisting of at least one or two syllable patterns, which is between one to four letters in the basic or root words.

2.2. Morphology

Morphology is a general technique that widely used in popular rule-based stemming by taking advantage of the affixes feature. Basically, the formation of affix in Indonesian and Sundanese language has several similarities. However, the Sundanese language has its own different rules in the formation of these affix words. Sundanese words are built by morphemes which are divided into three parts, namely free morphemes, semi-free morphemes, and *kauger* morphemes [17].

1. **Free morphemes:** morphemes that can stand alone in a sentences or morphemes that can directly become a basic word.
2. **Semi-free morphemes:** the half free morphemes that can have a lexical meaning and have a grammatical meaning.
3. ***Kauger* morphemes:** morphemes that can't stand on their own as words in a sentence, usually sticking to other morphemes.

2.3. Word Derivation

In general, word derivation in the Sundanese language came from affixes combination, repetitions, compound, and abbreviations until a basic morpheme is obtained. The word formation for derivation process is depicted in Figure 3.

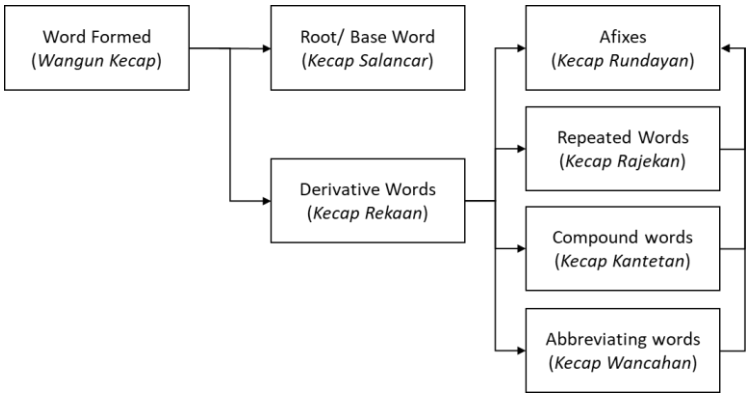

Figure 3. The Words Formation in Sundanese Morphology [17]

Based on the chart presented in Figure 3, the process of forming derived words involves four main processes, namely affixes, repeat words, compound words, and abbreviations.

2.4. Word Affixes

Sundanese language has rules in forming a word including affix, prolexeme, formative, and clitic. The rule of affixes used in the stemming process are rules that have been applied and can give good results. However, word formation in Sundanese does not only involve the use of affixes, but also applies other rules as described previously. In table 1 there are examples of word formations from these affixes which in Sundanese have many variations such as prefixes, infixes, suffixes and confixes.

Table 1. The Affixes form in Sundanese

Affixes	Function	Example
Prefix	pa-	*pa-tani* = 'farmer'
	pi-	*pi-wulang* = 'advice'

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Affixes	Function	Example
	sa-	*sa-imah* = 'at the same house'
	si-	*si-beungeut* = 'washing (face)'
	ti-	*ti-teuleum* = 'be drowned'
	(pa)ting-	*pating-gorowok* = 'shout (each other)'
	di-	*di-tulis* = 'written'
	ka-	*ka-tincak* = 'stepped on'
	mang-	*mang-rupa* = "form - (have the) shape of'
	ba-	*ba-layar* = 'sail'
	sang (nyang)-	*sang-hareup, nyang-hareup* = 'front -facing of'
	a-	*a-puputra* = 'have a children (son)'
	ma-	*ma-gawe* = 'plowing'
	bala-	*bala-harti* = 'misunderstanding'
	pari-	*pari-basa* = 'proverb'
	-ar-	*barudak* = 'children'
Infix	-al-	*la/ieur* = 'dizzy - dizziness'
	-um-	*gumeulis* = 'act as pretty'
	-in-	*sinerat* = 'written'
	-an	*cai-an* = 'water - irrigated'
	-eun	*cicing-eun* = 'silent - '
	-keun	*leutik-keun* = 'small - minimize (size)'
Suffix	-na	*imah-na* = 'his house - possessive pronoun'
	-ing	*wireh-ing* = 'therefore'
	-ning	*wantu-ning* = 'because of'
	-a	*mugi-a* = 'hope - hopefully'
	-i	*saksen-i* = 'witness - be a witness'
	pa- (m, n, ng, ny)	*pa-nanya* (*tanya*) = 'question'
	pa- … -an	*pa-mere* (*bere*)-*an* = 'awarding '
Affixes Combination (Confix)	mang- … -keun	*mang-mawa*(*bawa*)-*keun* = 'bring'
	sa- … -na	*sa-mahi-na* = 'enough'
	sa- … -keun	*di-beak-keun* = 'be finished'
	di- … -keun	*ka-asup-an* = 'entry'
	ka- … -an, -(a)- … -na	*ka-balik-an-na* = 'reverse (with possessive pronoun)'
	ka- .. -keun	*ka-dieu-keun* = 'over here'

In addition to the affixes in table 1, the Sundanese language has many forms of affixes called allomorphs that formed due to the affix N- which meets the initial consonant letter of the basic word, causing sound changes to become *m-*, *n-*, *ny-*, *ng-*, *nga-*, and *nge* [17] as list in table 2.

Table 2. The Allomorph Form in Sundanese

Affixes		Phoneme	Word Examples	
Morph	Allomorph		Root Words	Word Derivation
N-	m-	b	*baca*	*maca*
		p	*pacul*	*macul*
		f	*fitnah*	*mitnah*
	n-	t	*tanya*	*nanya*
			tonjok	*nonjok*
	ny-	c	*cokot*	*nyokot*
		s	*sugu*	*nyugu*
	ng-	k	*kawih*	*ngawih*
	nga-	b	*bedah*	*ngabedah*
		d	*dahar*	*ngadahar*
		g	*goler*	*ngagoler*
		h	*huit*	*ngahuit*
		j	*jentul*	*ngajentul*
		l	*lamar*	*ngalamar*
		m	*manah*	*ngamanah*
		ny	*nyatakeun*	*nganyatakeun*
		r	*rakit*	*ngarakit*
		w	*wajit*	*ngawajit*
		y	*yuga*	*ngayuga*
	nge		*bom*	*ngebom*

Affixes		Phoneme	Word Examples	
Morph	Allomorph		Root Words	Word Derivation
			cet	*ngecer*

Apart from the word -forming affixes in Table 1 and Table 2, this study includes prolexem in the form of prefix or suffix that is included with words and forms new words that are classified into three types, namely *panyebut*, *panglaku*, dan *pamilang* [17].

Table 3. The Prolexem form in Sundanese

Prolexem	Position		Example
	Front of word	End of word	
Panyebut	para- pari- -man pra- pasca-	-man -wan -wati	***para**guru*, ***pari**basa*, ***pra**duga*, ***pasca**panén*, *budi**man***, *warta**wan***, *seni**wati***
Panglaku	barang- pada- sili(h)- pili(h)		*barangbeuli*, *padadatang*, *silihsimbeuh*, *piligenti*
Pangwilang	éka- dwi- tri- catur- panca- dasa-		***éka**suku*, ***dwi**fungsi*, ***tri**darma*, ***catur**tunggal*, ***panca**sila*, ***dasa**sila*

2.3. Dataset

In this study, the articles data in Sundanese language was taken from http://majalah-balebat.blogspot.com/ which were also used in similar studies [14]. However, the article data from http://dewandakwahjabar.com/ had difficulty accessing it so we could not use the data source from that page. However, in this study only 69 articles were used because there was one article that had the same content. Furthermore, the data is broken down into single word which contains about 2945 words (containing affixes, allomorphs, and others) that will be used for the stemming testing process after removed the duplicate words.

In order to support accordance with the expected goals, the stemming process in this study also uses a corpus-based process which will be used as comparative data. The corpus data taken from Sundanese dictionary [19] which approximately 25988 words and divide into two categories which 18117 is with affixes used and 7871 is a single word used for the root word that use as a comparison data which depicted in the figure 4.

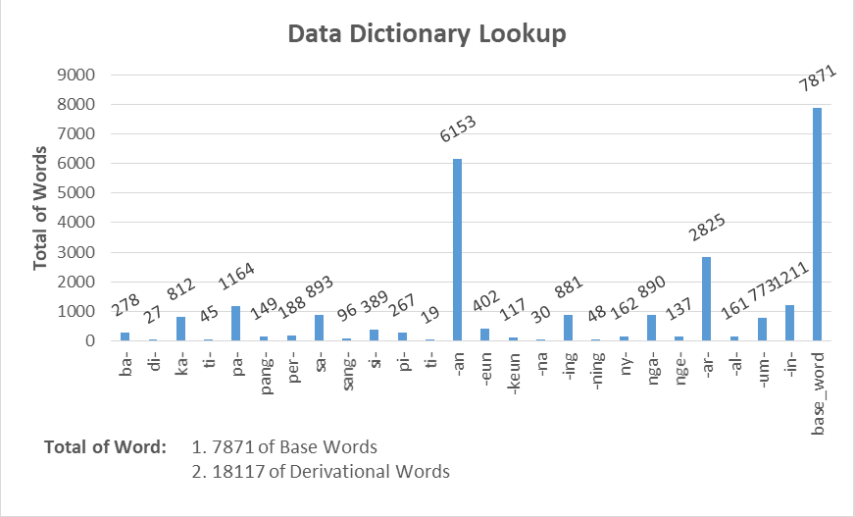

Figure 4. The Data Dictionary (Corpus)

p-ISSN: 2528-1682
e-ISSN: 2527-9165

3. RESULTS AND DISCUSSION

In general, the pattern of affixes in Sundanese is similar to other languages, which have prefixes, infix, and suffixes. But, in Sundanese there are more words that have a combination of affixes (confix) with the presence of allomorphs. This study uses a multi rule-based approach by utilizing morphological and phonology patterns, and the use of corpus data as a comparison of the stemming results from the rule-based. Therefore, the first step in the stemming process based on the morphology [5,6-9,12-15] rule will remove the affix first and also remove the pro-lexeme contained in the test data.

The next step is to adjust the words from the affix removal process result based on their syllable pattern [15]. As previously explained, there are five canonical forms of the Sundanese syllable pattern, namely, *ekasuku* (one of a syllable), *dwisuku* (two of a syllable), *trisuku* (thee of a syllable), *catursuku* (four of a syllable), a syllable), and *pancasuku* (five of a syllable). The removal of affixes will have an effect on the syllable pattern so that the execution process to obtain the basic word becomes easier. These patterns are convert the represent vowels and consonants according to the root word or basic word. A root word or basic word has a pattern that has at least one letter in the *ekasuku* and the maximum is 12 letters in the *pancasuku* [17]. In this case, the stemming algorithms using syllable patterns can give good results in returning a word to a basic or root word.

After getting the results of removing affixes and comparing the canonical forms, the next step is to compare these results with the corpus data. This step aims to avoid the over-stemming in the process of removing affixes or the word does not contain affixes even though they have the same canonical form. For example, table 4 presents a list of words that may doesn't have an affixes but have the same canonical form in a syllable pattern. Therefore, the use of corpus-based stemming becomes very important for languages that have special rules such as those found in Sundanese.

Table 4. Example Word of Stemming Problem

Kecap Rajekan	***Kecap Kantentan***	***Kecap Wancahan***	**Naming Object (people)**	**Loan Word**
ajleng-ajlengan	alang-ujurna	disbudpar	hardjadinata	implikasi
seuseungitan	lawanggintung	menko	Iskandardinata	internasional
titatarajong	ngaler-ngidul	mensesneg	kartadjumena	komputerisasi
undak-undakan	parungsinga	dékah	kartasasmita	kreativitas

In this study, a prototype of a web-based stemming application was made which was used for stemming experiments on test data that had been obtained previously. The stemming process is done by removing the affix first, then looking at the canonical form, and comparing it with the corpus data. The result of stemming is displayed on an interface which contains some information related to affix type, canonical form, and root word. Figure 5 below is an example of the appearance of the stemming process.

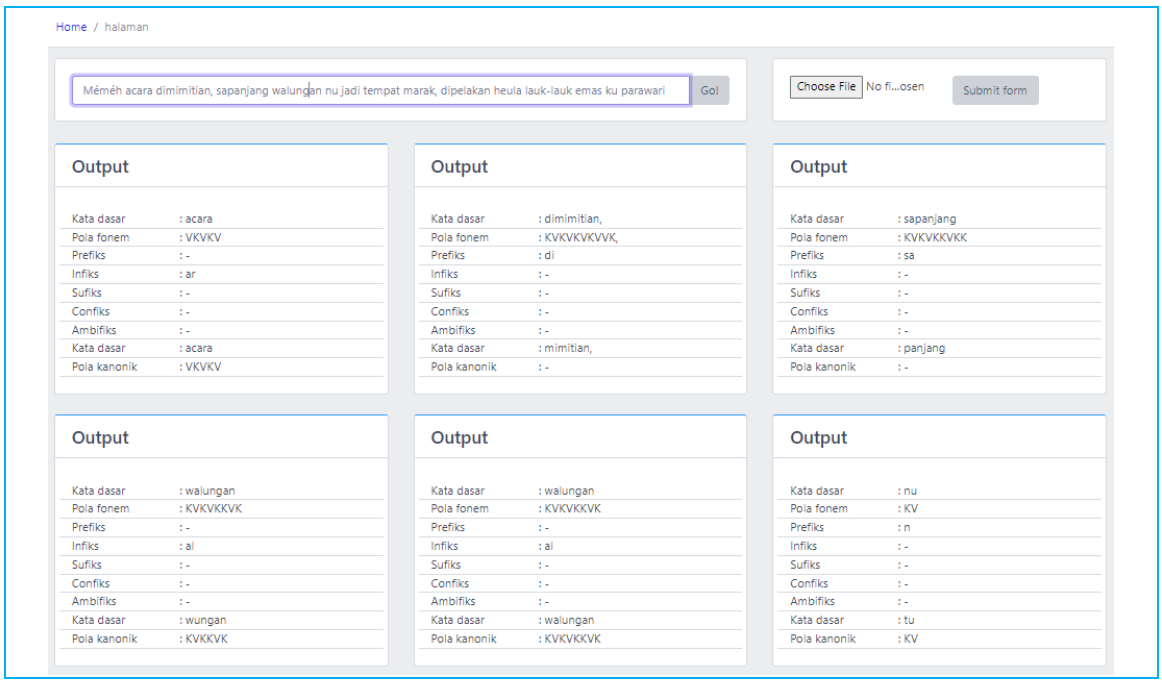

Figure 5. The Prototype Interface of Stemmer Application

In this study, the stemming process involves the rules of morphology, phonology, and data corpus, each of which has a function as described in the previous section. The results obtained from the execution process of 2945 data, which 2827 data were successfully executed with basic word output. Then, as many as 335 data can be executed and are divided into two categories, which 221 can be executed but not with the basic word output. While the rest is data that is displayed in its original form without going through the stemming process. It's because the words consist of loan words that have not been registered in the syllable pattern or insert into the corpus data. Thus, the accuracy of stemming results in this study reached 92.18% which the use of multi rule-base (morphology and phonology) and corpus-based technique can be used as a reference in the stemming process, especially in Sundanese. Figure 6 below illustrates the results of stemming processing using multi-rule-based (morphology and phonology) and corpus-based processing.

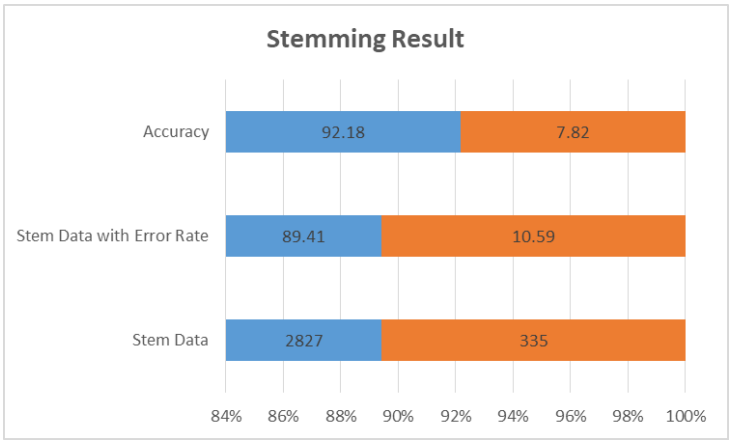

Figure 6. The Stemming Result

4. CONCLUSION

In this study, the stemming process involves several methods including morphological (with affix and pro-lexeme removal), syllable (canonical) pattern, and corpus data as a comparison of the final results of stemming. Affix removal is performed on words that have a prefix, infix, suffix, and confix which also the allomorph and pro-lexeme then compared with the syllable pattern and data dictionary from the provided corpus. The rule of affixes removal can remove the corresponding affixes and derived to the simplest word (base word). The rule of syllable can compares two words that have a syllable pattern and execute the basic words quickly according to their syllable. Finally, the use of corpus will compare between stemming words

and words that already exist in the data dictionary so that they can improved the accuracy and reduce the over-stemming problems that occur in the stemming process. Thus, the accuracy of stemming results in this study reached 92.18% which the use of multi rule-base (morphology and phonology) and corpus-based technique can be used as a reference in the stemming process, especially in Sundanese. Although this study has resulted in a fairly high stemming accuracy value, it still needs to be improved so that the stemming process becomes more properly. For the future research, the stemming process can be adding rules and should be able to handle *kecap rajekan* (compound word) for examples the words *bapa-bapa*, *tujang-tajong*, and *plak-plik-pluk*, and *kecap wancahan* (abbreviation word) such as the example of the word dékah stands for *déwék mah* and kirata stands for *dikira-kira sugan nyata*.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Education, Culture, Research, and Technology who gave the opportunity for us to do this research with parent contract number 065 /SP2H/LT/DRPM/2021 and derivative contract number 069/SP2H/RDPKR-MONO/LL4/2021.

REFERENCES

[1] P. Willett, "The Porter stemming algorithm: then and now," Program, vol. 40, no. 3, pp. 219–223, Jul. 2006, doi: 10.1108/00330330610681295.
[2] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian," ACM Trans. Asian Lang. Inf. Process., vol. 6, no. 4, pp. 1–33, Dec. 2007, doi: 10.1145/1316457.1316459.
[3] A. Purwarianti, "A non deterministic Indonesian stemmer," Proc. 2011 Int. Conf. Electr. Eng. Informatics, ICEEI 2011, no. October, 2011, doi: 10.1109/ICEEI.2011.6021829.
[4] A. A. Damar, K. Dewi, and U. M. Siti, "Penerapan Algoritma Paice atau Husk untuk Stemming pada Kamus Bahasa Inggris ke Bahasa Indonesia," J. Tek. Inform., vol. 6, no. 2, Oct. 2013, doi: 10.15408/jti.v6i2.2031.
[5] A. S. Rizki, A. Tjahyanto, and R. Trialih, "Comparison of stemming algorithms and its effect on Indonesian text processing," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 17, no. 1, p. 95, Feb. 2019, doi: 10.12928/telkomnika.v17i1.10183.
[6] Y. Anistyasari and E. Hariadi, "Algoritma Baru Pembentukan Kata Dasar Pada Proses Stemming Bahasa Indonesia," Pros. SNRT (Seminar Nas. Ris. Ter., vol. 5662, no. November, pp. 70–76, 2019.
[7] F. Amin and Purwatiningtyas, "Stemmer Bahasa Jawa Ngoko dengan Metode Affix Removal Stemmer (Rule Base Approach)," J. Teknol. Inf. Din., vol. 21, no. 1, pp. 16–24, 2016.
[8] N. Hidayatullah, A. P. Wibawa, and H. A. Rosyid, "Penerapan ECS Stemmer untuk Modifikasi Nazief & Adriani Berbahasa Jawa," vol. 3, no. 3, pp. 343–348, 2019.
[9] R. Maulidi, "Stemmer Untuk Bahasa Madura Dengan Modifikasi Metode Enhanced Confix Stripping Stemmer," in Prosiding Seminar Nasional FDI 2016, 2016, no. December, pp. 12–15.
[10] G. Ngurah, M. Nata, and P. P. Yudiastra, "Stemming teks sor-singgih Bahasa Bali," Konf. Nas. Sist. Inform. 2017 STMIK, no. Agustus, pp. 608–612, 2017.
[11] M. Agus, P. Subali, C. Fatichah, and D. Informatika, "Kombinasi Metode Rule-Based Dan N-Gram Stemming Untuk Mengenali Stemmer Bahasa Bali," J. Teknol. Inf. dan Ilmu Komput., vol. 6, no. 2, 2019, doi: 10.25126/jtiik.201961105.
[12] D. Junaedi, O. Herlistiono, and D. Akbar, "Stemmer For Basa Sunda," pp. 275–278, 2010.
[13] A. Purwoko, "Model Stemming Berbasis kamus untuk dokumen berbahasa sunda," INSTITUT PERTANIAN BOGOR, 2011.
[14] A. A. Suryani, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "The Rule-Based Sundanese Stemmer," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 17, no. 4, pp. 1–28, Aug. 2018, doi: 10.1145/3195634.
[15] A. Sutedi, R. Elsen, and M. R. Nashrulloh, "Sundanese Stemming using Syllable Pattern," vol. 6, no. 2, pp. 218–224, 2021, doi: 10.15575/join.v6i2.812.
[16] I. Baidillah et al., Direktori Aksara Sunda untuk Unicode, 1st ed. Dinas Pendidikan Provinsi Jawa Barat, 2008.
[17] D. Sudaryat, Yayat, A. Prawirasumantri, and K. Yudibrata, Tata Basa Sunda Kiwari. Bandung: Yrama Widya, 2013.
[18] L. S. Faznur et al., "Komparasi fonem bahasa sunda dan bahasa indonesia dalam buku teks," Pena Literasi J. Pendidik. Bhs. dan Sastra Indones., vol. 2, no. 2, pp. 105–114, 2019.
[19] R. A. Danadibrata, Kamus Basa Sunda, 4th ed. Bandung: Panitia Penerbitan Kamus Basa Sunda dan PT. Kiblat Buku Utama, 2015.