

# Sundanese Stemming using Syllable Pattern

**Ade Sutedi<sup>1</sup>, Rickard Elsen<sup>2</sup>, Muhammad Rikza Nashrulloh<sup>3</sup>**

<sup>1,2,3</sup>Department of Computer Science, Institut Teknologi Garut, Indonesia

---

## Article Info

### *Article history:*

Received September 15, 2021

Revised October 18, 2021

Accepted November 09, 2021

Published December 22, 2021

---

### **Keywords:**

Phonology

Stemming

Sundanese

Syllable

---

## ABSTRACT

Stemming is a technique to revert the word derivation to the root or base word. Stemming is widely used for data processing such as searching word indexes, translating, and information retrieval from a document in the database. In general, stemming uses a morphological pattern from a derived word to produce the original or root word. In the previous research, this technique faced over-stemming and under-stemming problems. This study aims to address these issues by enhancing the stemming process through the use of syllable patterns based on the phonological rules of Sundanese. The stemming result for syllable patterns gets an accuracy of 89% and the execution of the test data resulted in 95% from all the basic words. This research found that this simple algorithm has the advantage of adjusting the position of the syllable pattern with the target word. Due to some data shortage constraints (typo, loan-word, non-deterministic word with syllable pattern), enhancing accuracy is achievable through methods such as word adjustment and the incorporation of reference dictionaries. In addition, this algorithm has a drawback that causes the execution to be over-stemming.

---

### *Corresponding Author:*

Ade Sutedi,

Department of Computer Science,

Institut Teknologi Garut,

Jl. Mayor Syamsu No. 1 Garut, Indonesia

Email: adesutedi@itg.ac.id

---

## 1. INTRODUCTION

There are variations in speech and written language the Sundanese language in several regions, but the Sundanese Priangan dialect has become the standard for Sundanese society, especially in West Java [1]. However, the data resources and studies such as text mining, machine learning, and natural language processing for Sundanese are limited. It is because the influence of foreign languages and Indonesian as the official language has strengthened [2]. Therefore, this research is driven by the lack of corpus data availability and development activities in software tools used for various applications, especially those related to the Sundanese language. Notably, recent studies that discuss stemming techniques for regional languages have begun such as Javanese [3], [4], Madura [5], Balinese [6], [7], Batak [8], and also Sundanese [9], [10], [11].

The stemming technique in previous research was developed in morphology form including derivation and inflectional affixes removal for regional language [3], [4], [5], [6], [7], [9], [10], [11], Indonesia [12], [13], [14], [15], and English [16], [17]. Several stemming methods used in previous research are rule-based [3][10][13], the combination of rule-based with corpus or dictionary [7], [11], also rule-based with n-gram [7]. The rule-based technique adopted from Porter's algorithm focuses on the process of removing the suffix to the right of the root word [16]. This method is widely applied to stemming development for Indonesian and later adapted to stemming for regional languages. Another technique was developed which applied to several regional languages stemming using confix stripping and enhanced for confix stripping (ECS) method. Although the ECS technique has good stemming results for Indonesian [3], but not for regional languages.

Meanwhile, stemming for the Sundanese language becomes difficult when it covers allomorph which could be a kind of base word or original word [9], [11]. In addition, the difference of morphological patterns in regional languages has many differences compared with Indonesian or English language such as derived words constructed by the combination of prefixes, infixes, suffixes, or confixes [10], [11]. So, the application

of the rule-based method becomes ineffective in some cases especially if it depends on the completeness of the dictionary database because there are even some words that have no equivalent in Indonesian [18]. To overcome this problem, the research uses a different rule-based approach by utilizing syllable patterns that form derived words in the Sundanese language. This process is expected to be able to improve several cases that have not been resolved in previous studies.

## 2. METHOD

The study of the sound of language becomes the basis of writing or grammar. Phonemes in Sundanese can have a variety of pronunciations corresponding to their place in the noun or word, typically not distinguishing the meaning [18]. There are twenty-five phonemes in Sundanese [18], [19], [20], which is divided into two categories namely vowel (Swara) and consonant or append letter (Wianjana). For the 7 vowel distribution i.e.: /a/ [a], /i/ [i], /u/ [u], /é / [ɛ], /o/ [o], /eu/ [ö], /e/ [e] and 18 consonants, i.e.: /b/ [b], /c/ [c], /d/ [d], /g/ [g], /h/ [h], /j/ [j], /k/ [k], /l/ [l], /m/ [m], /n/ [n], /ny/ [ñ], /ng/ [ŋ], /p/ [p], /r/ [r], /s/ [s], /t/ [t], /w/ [w], and /y/ [y]. Along with the development of national and foreign language uses, there are additional consonants, i.e.: /f/ [f], /q/ [q], /v/ [v] , /x/ [x], and /z/ [z] [21].

### 2.1. Research Flow

In this research, the stemming process using a syllable pattern algorithm was developed based on the following research flow depicted in Figure 1.

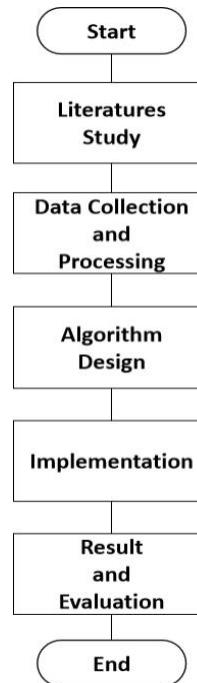


Figure 1. Research Flow for Stemming using Syllable Pattern

### 2.2. Syllable Pattern

Syllables are the basic word-forming components (morphemes) in the Sundanese language which generally consist of two syllables. Generally, the Sundanese language consists of vowels and consonant arrangements as a system of sound in the form of syllables or canonic. The canonical form of the word in the Sundanese language is as follows [22]:

- V - Vowel
- CV - Consonant Vowel
- VC - Vowel Consonant
- CVC - Consonant Vowel Consonant
- CCV - Consonant Consonant Vowel
- CCVC - Consonant Consonant Vowel Consonant

In addition, due to the influence of the use of national and foreign languages, there are several canonical patterns as follows:

**VCC** - Vowel Consonant Consonant

**CVCC** - Vowel Consonant Consonant

**CCCV** - Consonant Consonant Consonant Vowel

**CCVCC** - Consonant Consonant Vowel Consonant Consonant

**CCCVC** - Consonant Consonant Consonant Vowel Consonant

**CVC** - Consonant Vowel Consonant Consonant

For all Sundanese syllable pattern, the writing rules are as follows [16]:

1. Sundanese words consist of *Ekasuku* (one of the syllables), *Dwisuku* (two of syllables), *Trisuku* (three of syllables), *Catarsuku* (four of syllables), and *Pancasuku* (five of syllables).

Table 1. Canonical Form in Sundanese

<i>Ekasuku</i>	<i>Dwisuku</i>	<i>Trisuku</i>	<i>Catarsuku</i>	<i>Pancasuku</i>
V	V-V	V-V-CV	V-CV-CV-CV	V-CV-CV-CV-CVC
VC	V-CV	V-CV-CV	V-CV-CV-VC	CV-CV-CV-CV-CV
CV	V-CVC	V-CV-CVC	V-CVC-CV-VC	CV-CV-CV-CVC-CV
CVC	V-CCV	VC-CV-CV	V-CV-CV-CVC	CV-CV-CV-CVC-CVC
CCV	V-CCVC	VC-CV-VC	VC-CV-CV-CVC	CVC-CV-CV-CV-CVC
CCVC	VC-VC	VC-CV-CVC	VC-CV-CV-VC	CV-CV-CV-CVC-CVC
	VC-CV	VC-CV-CCVC	CV-CV-CV-CV	CV-CV-CV-CV-VC
	VC-CCV	VC-CV-CCV	CV-CV-CV-CVC	
	VC-CCVC	CV-CV-VC	CV-CV-CV-CV	
	CV-V	CV-CV-V	CV-CV-CVC-CVC	
	CV-VC	CV-CV-CVC	CV-CV-VC-CVC	
	CV-CV	CV-CVC-CVC	CV-CV-V-CV	
	CV-CVC	CV-V-CVC	CVC-CV-CV-CVC	
	CV-CCV	CV-CVC-CCVC	CV-CVC-CV-CV	
	CV-CCVC	CVC-CVC-VC	CV-CV-V-CVC	
	CVC-CV	CVC-CV-CV	CVC-CV-CV-CV	
	CVC-CVC	CV-CVC-CV	CVC-CV-CV-V	
	CVC-CCV	CVC-CV-CVC	CVC-CV-V-CVC	
	CVC-CCVC	CVC-CV-VC	CV-CVC-CV-CVC	
		CVC-CVC-CVC	CV-CVC-CV-VC	
		CCV-CV-CVC		

2. Consonants do not stand alone as syllables,
3. Vowels can stand alone as syllables,
4. Consonant groups can only be at the beginning of the,
5. The second consonant in the consonant cluster generally consists of a few l, r, and y consonants,
6. *Ekasuku* phoneme arrangement is the basic pattern of two or more syllables.

### 2.3. Syllable Distortion

There are different ways to distort the original words (root words) and the affixes word morphemes.

#### a. The original word distortion

1. If in the middle of a word there are two vowels, the way of separating them should be between those two vowels.

*da-un*                    <*daun*>  
*leu-it*                    <*leuit*>

2. If in the middle of a word, there is a consonant between two vowels, the way to separation must be before the consonant.

*a-di*                    <*adi*>  
*sa-reng*                    <*sareng*>  
*wa-yang*                    <*wayang*>

3. If in the middle of a word there is a consonant denoted by two letters, the letter renditions are not painted. the way to separation is before or after the letter cluster.

*di-nya*                    <*dinya*>  
*te-nga*                    <*tengah*>

4. If in the middle of a word two consonants are not cluster (*rendon*) consonants, a way of separating between the two consonants.

<i>ang-klung</i>	<i>&lt;angklung&gt;</i>
<i>sar-ta</i>	<i>&lt;sarta&gt;</i>
<i>tek-tek</i>	<i>&lt;tékték&gt;</i>

5. If in the middle of a word there are two or more consonants is a *wianjana* mound, the consonant rendon is not painted. The disfigure before the wianjana mound.

<i>am-prok</i>	<i>&lt;amprok&gt;</i>
<i>ga-plék</i>	<i>&lt;gaplek&gt;</i>
<i>nam-bru</i>	<i>&lt;nambru&gt;</i>

**b. The affixes word distortion**

1. Prefixes, suffixes, and regular particles are written tied to the original word, the way of separation must be independent.

<i>di-ba-wa</i>	<i>&lt;dibawa&gt;</i>
<i>da-har-eun</i>	<i>&lt;dahareun&gt;</i>
<i>da-han-na</i>	<i>&lt;dahanna&gt;</i>

2. Prefixes, suffixes, and particles underwent a change of written form as follows.

<i>pa-na-nya</i>	<i>&lt;pananya&gt;</i>
<i>pa-nga-la</i>	<i>&lt;pangala&gt;</i>
<i>da-har-eun-a-na</i>	<i>&lt;dahareunana&gt;</i>

3. Infix of the written note with the word it comes from, distorting like distorting the original word.

<i>ba-ra-la</i>	<i>&lt;barala&gt;</i>
<i>da-la-har</i>	<i>&lt;dalahar&gt;</i>
<i>gu-meu-lis</i>	<i>&lt;gumeulis&gt;</i>
<i>pi-nang-gih</i>	<i>&lt;pinanggih&gt;</i>

4. Infix that changes its place in front of the word originally, is considered equal to the next rank.

<i>ar-u-lin</i>	<i>&lt;arulin&gt;</i>
<i>al-u-dur</i>	<i>&lt;aludur&gt;</i>
<i>ra-cleng</i>	<i>&lt;racleng&gt;</i>
<i>um-a-mis</i>	<i>&lt;umamis&gt;</i>

#### 2.4. Corpus Data

In this research, the corpus data was taken from which contains Sundanese wave sound and a TSV file transcript [23]. We utilized sections for male and female data in the TSV file, which we then combined it into a single text file. This file contains 4,213 sentences and 61,383 tokens of words with duplication. For the input process, we used the words after applying a filtering process. Consequently, 3,643 unique words were selected as testing data. Then, we manually validated for every word and we had 494 words with affixes form. This data served as the basis for algorithm implementation.

#### 2.4. Design of Algorithm

The stemming design process in this study was conducted by developing a technique for forming affixes based on [10] with a combination of affixes [14] as shown in Figure 2.

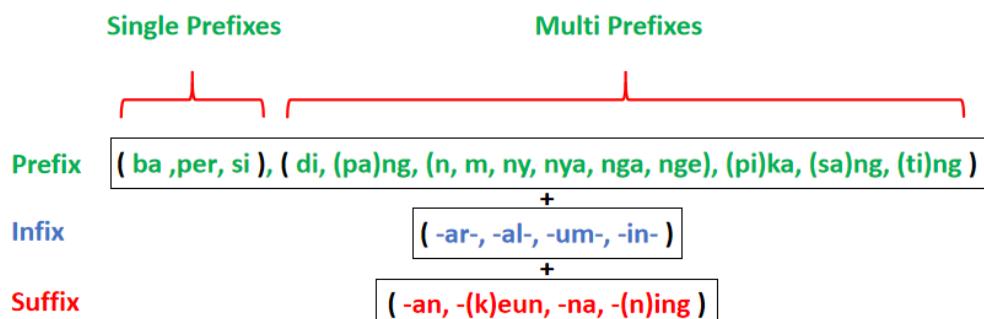


Figure 2. Affixes combination in Sundanese

Based on the rules of syllables and word derivatives in Sundanese, the design of the stemming algorithm with a syllable pattern can be seen in Figure 3.

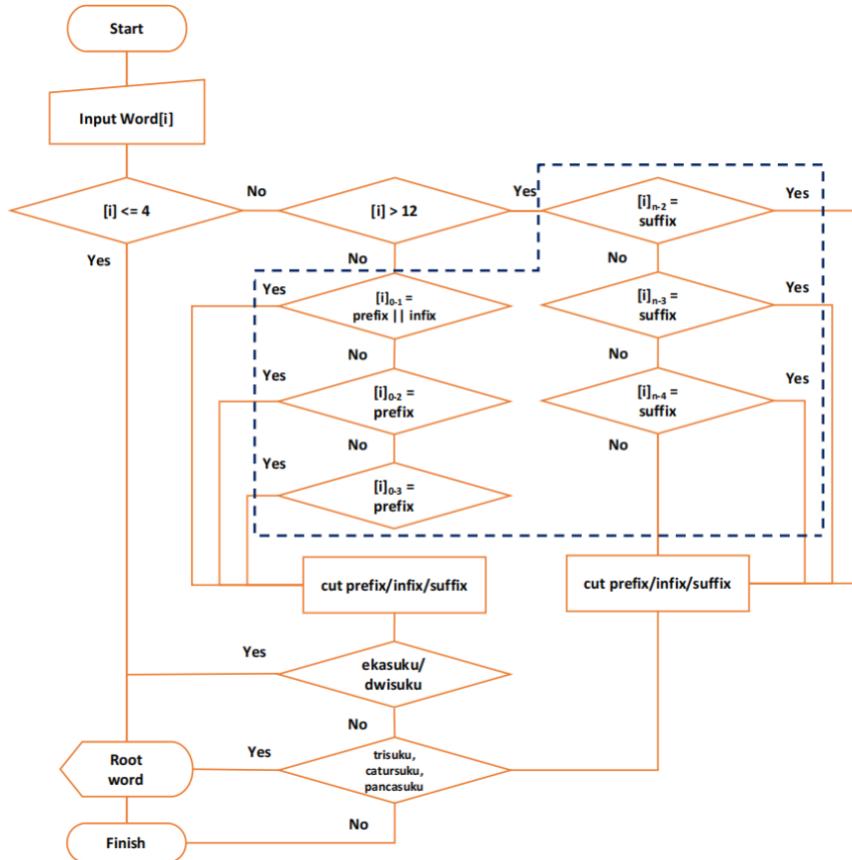


Figure 3. Stemming algorithm with syllable pattern

### 3. RESULTS AND DISCUSSION

After completing the research stages, in general, the use of algorithms with the syllable pattern feature can be applied to the stemming process, especially for the Sundanese language. We present the results and discussion of this study in the subsequent sections.

#### 3.1. Result

In this study, the process of testing stemming algorithms using syllable patterns was performed on data sourced from 4,213 sentences, 61,383 tokens of words with duplication [23]. Then filtered it with a result of 3,643 unique words were obtained with a total of 494 affixes words. stemming results were obtained 440 syllable patterns (canonical) and 469 base words. Thus the stemming algorithm result using this syllable pattern has an accuracy of 89% with a base word result approach of 95%.

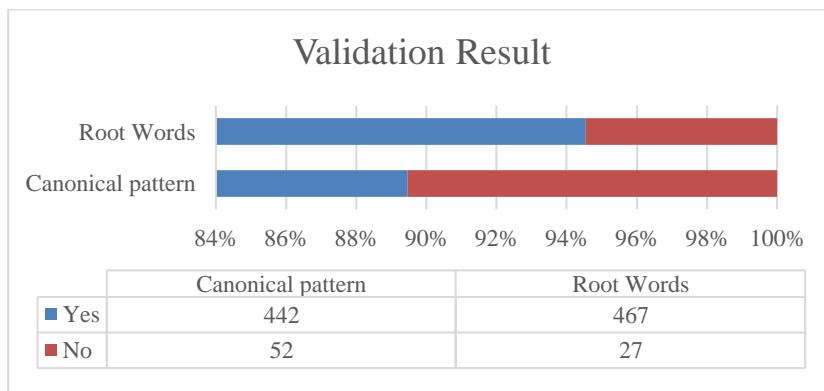


Figure 4. Stemming Process Result

### 3.2. Discussion

Syllable patterns have an important role in the word derivation in Sundanese. Sundanese syllables consist of *Ekasuku* (one of the syllables), *Dwisuku* (two of syllables), *Trisuku* (three of syllables), *Catusuku* (four of syllables), and *Pancasuku* (five of syllables). These patterns are conversions that represent vowels and consonants as well as become a root or basic word in a word. A root or basic word has a pattern that has at least one letter in the *Ekasuku* and a maximum of 12 letters in the five letters. Thus, the application of the stemming algorithm in this study is divided into three main processes following the syllable pattern. The canonical comparison process of a word with its syllables is illustrated in Figure 5.



Figure 5. The canonical form of the word

Stemming algorithms using syllable patterns can give good results in returning a word to a root word or root word. However, because Sundanese grammar is different from Indonesian, English, and others. then, several cases arise and become obstacles when the stemming process is run. First, the use of the Sundanese alphabet has differences, for example "/eu/ [ö]", "/ny/ [ñ]", and "/ng/ [ŋ]" which can lead to confusion in the syllable conversion process [18]. Second, the use of loan-word from Indonesian appears in the dataset [23], including for example "*Administratifna*", "*Efektifna*", "*Kesenian*", "*Perumahan*", and "*Perpustakaan*" which causes the stemming process unable to produce appropriate canonical patterns nor root words. Third, the use of nasal prefixes (allomorph) in Sundanese cannot be distinguished because the canonical form of syllables can have many possible words (non-deterministic) so an additional dictionary is needed as a reference for the proper use of nasal prefixes [13].

## 4. CONCLUSION

In this study, stemming using syllable patterns can be applied to return affixed words into the root or base word with an accuracy of 89%. The results of the execution of the test data resulted in 95% of the basic words. This simple algorithm has the advantage of adjusting the position of the syllable pattern with the word being stemmed. However, certain limitations, such as data scarcity (including typos, loan-words, and non-deterministic syllable patterns) impede the achievement of a higher accuration rate. By refining word adjustments and integrating reference dictionaries, we anticipate enhancing accuracy by 3-5%. In addition, this algorithm has a drawback that causes the execution to be over-stemming. Further research could focus on several areas to enhance the accuracy and efficiency of the stemming algorithm and paving the way for its broader adoption in natural language processing applications.

## ACKNOWLEDGEMENTS

The authors are grateful for the research funded and supported by the Ministry of Education, Culture, Research, and Technology with parent contract number 065 /SP2H/LT/DRPM/2021 and derivative contract number 069/SP2H/RDPKR-MONO/LL4/2021. Institut Teknologi Garut with contract number 3/STTG/LPPM/SP-LPPM/II/2021.

## 5. REFERENCES

- [1] E. Z. Arifin, "Bahasa sunda dialek priangan," *Pujangga: Jurnal Bahasa dan Sastra*, vol. 2, no. 1, pp. 1–44, 2016.
- [2] R. U. Fadila, "Bahasa Sunda Hadapi Tantangan Besar, Pemerintah Lakukan Beragam Upaya," *Pikiran Rakyat*. Accessed: Jan. 01, 2021. [Online]. Available: <https://www.pikiran-rakyat.com/pendidikan/pr-01342765/bahasa-sunda-hadapi-tantangan-besar-pemerintah-lakukan-beragam-upaya>
- [3] N. Hidayatullah, Aji Prasetya Wibawa, and Harits Ar Rosyid, "Penerapan ECS Stemmer untuk Modifikasi Nazief & Adriani Berbahasa Jawa," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 3, pp. 343–348, Dec. 2019, doi: 10.29207/resti.v3i3.994.

[4] F. Amin and P. Purwatiningtyas, "Stemmer Bahasa Jawa Ngoko dengan Metode Affix Removal Stemmers (Rule Based Approach)," *Dinamik*, vol. 21, no. 1, pp. 16–24, 2016.

[5] R. Maulidi, "Stemmer untuk Bahasa Madura dengan Modifikasi Metode Enhanced Confix Stripping Stemmer," in *Prosiding Seminar Nasional FDI*, 2016, pp. 12–15.

[6] G. N. M. Nata and P. P. Yudistastra, "Stemming teks sor-singgih Bahasa Bali," *E-Proceedings KNS&ISTIKOM Bali*, pp. 608–612, 2017.

[7] M. Agus, P. Subali, C. Fatichah, and D. Informatika, "Kombinasi Metode Rule-Based Dan N-Gram Stemming Untuk a Combination of Methods Rule-Based and N-Gram Stemming To Recognize Balinese Language Stemmer," *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIK)*, vol. 6, no. 2, 2019.

[8] M. A. Muchtar *et al.*, "Separation of Basic Words in Angkola Batak Text Documents using Enhanced Confix Stripping Stemmer Case: Mandailing Ethnic," *IOP Conf Ser Mater Sci Eng*, vol. 648, no. 1, p. 012024, Oct. 2019, doi: 10.1088/1757-899X/648/1/012024.

[9] D. Junaedi, I. O. Herlistiono, and D. Akbar, "Stemmer for 'Basa Sunda,'" Seminar Nasional Ilmu Komputer Universitas Dipenogoro, Teknik Informatika ..., 2010.

[10] A. Purwoko, "Model Stemming Berbasis kamus untuk dokumen berbahasa sunda," *INSTITUT PERTANIAN BOGOR*, 2011.

[11] A. A. Suryani, D. H. Widayantoro, A. Purwarianti, and Y. Sudaryat, "The Rule-Based Sundanese Stemmer," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 4, pp. 1–28, Dec. 2018, doi: 10.1145/3195634.

[12] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian : A confix-stripping Approach," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 6, no. 4, pp. 1–33, 2007, doi: 10.1145/1316457.1316459.

[13] A. Purwarianti, "A non deterministic Indonesian stemmer," in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, IEEE, Jul. 2011, pp. 1–5. doi: 10.1109/ICEEI.2011.6021829.

[14] D. S. Maylawati, W. B. Zulfikar, C. Slamet, and M. A. Ramdhani, "An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media," in *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*, 2018.

[15] 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, Mar. 2021, doi: 10.1088/1757-899X/1098/3/032044.

[16] P. Willett, "The Porter stemming algorithm: then and now," *Program*, vol. 40, no. 3, pp. 219–223, Jul. 2006, doi: 10.1108/00330330610681295.

[17] V. Gurusamy and S. Kannan, "Performance Analysis: Stemming Algorithm for the English Language," *International Journal for Scientific Research and Development*, vol. 5, pp. 2321–2613, 2017.

[18] L. S. Faznur and D. Nurhamidah, "Komparasi Fonem Bahasa Sunda Dan Bahasa Indonesia Dalam Buku Teks," *Pena Literasi*, vol. 2, no. 2, p. 105, May 2020, doi: 10.24853/pl.2.2.105-114.

[19] M. Sumantri, *Kamus Sunda-Indonesia*. 1985.

[20] B. R. Tamsyah, "Kamus Lengkap Sunda-Indonesia, Indonesia-Sunda, Sunda-Sunda," (*No Title*), 2003.

[21] I. Baidillah *et al.*, *Direktori Aksara Sunda untuk Unicode*, 1st ed. Dinas Pendidikan Provinsi Jawa Barat, 2008.

[22] Y. Sudaryat, A. Prawirasumantri, and K. Yudibrata, "Tata basa Sunda kiwari," (*No Title*), 2007.

[23] K. Sodimana *et al.*, "A Step-by-Step Process for Building TTS Voices Using Open Source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," in *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, ISCA: ISCA, Aug. 2018, pp. 66–70. doi: 10.21437/SLTU.2018-14.