Rule-based Dialect of Tulang Bawang Stemmer

1st Zaenal Abidin
Doctoral Program of Mathematics and Natural Sciences
Universitas Lampung
Bandar Lampung, Indonesia
2237061006@students.unila.ac.id

2nd Akmal Junaidi
Department of Computer Science
Universitas Lampung
Bandar Lampung, Indonesia
akmal.junaidi@fmipa.unila.ac.id

3rd Wamiliana
Department of Mathematics
Universitas Lampung
Bandar Lampung, Indonesia
wamiliana.1963@fmipa.unila.ac.id

4th Favorisen Rosyking Lumbanraja
Department of Computer Science
Universitas Lampung
Bandar Lampung, Indonesia
favorisen.lumbanraja@fmipa.unila.ac.id

5th Dian Kurniasari
Department of Mathematics
Universitas Lampung
Bandar Lampung, Indonesia
dian.kurniasari@fmipa.unila.ac.id

6th Rohmat Indra Borman
Faculty of Engineering and Computer Science
Universitas Teknokrat Indonesia
Bandar Lampung, Indonesia
rohmat_indra@teknokrat.ac.id

2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS) | 979-8-3315-1332-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICADEIS65852.2025.10933405

***Abstract*—Stemming, an essential procedure in natural language processing (NLP), diminishes words to their base forms, facilitating tasks such as information retrieval and sentiment analysis. Although stemming techniques for high-resource languages are well-developed, numerous low-resource languages, including dialect of Tulang Bawang, suffer from inadequate solutions owing to a scarcity of linguistic data and resources. Existing systems, including rule-based stemmers, have demonstrated efficacy in processing low-resource languages such as Indonesian and Javanese by utilizing established morphological rules. Nonetheless, these methods encounter considerable obstacles, such as restricted adaptability, inability to accommodate unusual root structures, and excessive dependence on fixed rules that might result in over- or under-stemming. Rule-based methodologies frequently misidentify roots when faced with intricate affixes or unconventional word forms. We introduce an improved rule-based Tulang Bawang Stemmer aimed at overcoming these constraints by enhancing current linguistic rules and integrating new patterns specific to the language's morphology. Assessed on 500 test samples and 200 independent test samples, our improved stemmer attained gold standard evaluation metrics of 96.2% and 94%, respectively, surpassing prior implementations in both precision and generalization. The findings demonstrate the potential of enhanced rule-based techniques to improving NLP for low-resource languages. Improved stemming performance enables better downstream applications, promotes more efficient text analysis, and advances research in underrepresented languages.**

Keywords— low-resource NLP, morphological processing, rule-based stemmer, tulang bawang stemming.

I. Introduction

Tulang Bawang (TB) is a dialect of the Lampung language utilized by populations in the Tulang Bawang area of Lampung Province [1]. It is part of the Lampung dialect of *nyo* language group. This dialect is distinguished by its distinctive vocabulary, pronunciation, and syntactic organization. Its utilization is predominantly observed in rural regions, while the impact of Indonesian is progressively apparent among younger demographics [2]. Research on the Tulang Bawang dialect is scarce, presenting prospects for additional investigation, especially in linguistics and language technology, including natural language processing (NLP). Besides Tulang Bawang, there is also the Abung dialect and the Pesisir dialect.

Natural language processing, evaluation and production in text and speech are possible for machines [3], [4]. The interface between computers and human language is the focus of Natural Language Processing (NLP) [3], [4]. NLP is widely used in machine translation, information retrieval, and sentiment analysis [3], [4]. Indonesian machine translation, information retrieval, and sentiment analysis research is significantly growing. Similar studies on local languages are scarce, leaving a research gap. Indonesia, with over 700 native languages, presents possibilities for NLP study [5]. Even though local languages are classified as low-resource languages, some even consider them underrepresented, they have the potential for NLP research [5], [6], [7], [8]. So machine translation, information retrieval, and sentiment analysis can potentially be done on local languages as well. Machine translation, information retrieval, and sentiment analysis research depend on text pre-processing [3], [4]. Natural language processing text preparation is based on stemming and lemmatization. Stemming removes affixes, returning words to their roots, whereas lemmatization examines morphology to determine dictionary form. These methods normalize text, which enhances linguistic data processing and comparability [9], [10]. Stemming algorithms such as Porter Stemmer and Snowball Stemmer are available for a variety of languages [9], [10], [11]. Many languages employ stemming and lemmatization. Stemming in Italic, Slavic, Uralic, Germanic, Indo-Aryan, Dravidian, Iranian, and Semitic languages demonstrates the importance of dealing with morphological complexity [9], [10].

Stemming research has been conducted in a number of Indonesian regional languages. The stemming methods used in some regional languages were Ruled-based, Brute Force or Table look up, Nazief-Adriani, Confix-Stripping, Enhanced Confix-Stripping, N-Gram Stemming, Syllable pattern, and Corpus-based. The regional languages used in stemming research start from [12], [13], [14] [15], Batak Angkola [16], Tetun [17], Minangkabau [18], [19], Rejang [20], [21], Javanese [22], Balinese [23], [24], and Madurese [25]. Each of these regional languages used rule-based or morphological techniques, and corpus-based while Lampung language used a brute force approach [26].

This TB dialect rule-based stemmer research differs from past research which employed brute force or table lookups. In prior investigations, stemming was accomplished solely by matching the word to be stemmed with a list of terms containing the base word. Previous research has the following flaws: (1) it cannot handle stemming if it encounters a word that is not in the list of attached words, and (2) the brute force or table look up approach does not employ morphological data [26].

Rule-based stemmer research addresses these shortcomings. Tulang Bawang is spoken in Tulang Bawang

regency, Lampung. The Tulang Bawang ethnic group uses it most. This language has complex prefixes, suffixes, and reduplication. Tulang Bawang is underrepresented in NLP research despite its cultural value. Insufficient language resources and tools hinder text processing and analysis programs. Due to its complex morphology and lack of standardization, stemming this language is difficult. Regional language stemming methods in Indonesia vary in efficacy. Regional language stemmers are usually language-specific. These limits require a more advanced stemming strategy for low-resource languages like Tulang Bawang.

II. Morphology of Tulang Bawang Dialect

Linguistic morphology studies word structure, form, and creation. The simplest units of language, morphemes such roots, prefixes, suffixes, infixes, confixes and reduplication, are examined [1], [2], [27]. Morphological studies also examine how morphemes combine to produce new words or change meanings. Morphology lets us study word production patterns across languages and how word structure affects syntax and semantics. Word structure and construction distinguish Tulang Bawang's morphology another dialect. This dialect uses prefixes, suffixes, infixes, confixes and reduplication to create words or change their meanings. All TB morphological explanations are from book [2] and a publication related to the morphology of the *Tulang Bawang* dialect was written by Farida Ariyani [27].

A. Base Verbs

Base verbs come from root words without affixation, reduplication, or composition. Basic *Tulang Bawang* verbs explain simple actions, attitudes, and activities with direct meaning. Simple actions like "*mengan*" (means “eat”) and "*cekak*" (means “go up”) are examples. The meanings and grammatical contexts of derived verbs are changed by adding prefixes or suffixes to these fundamental verbs.

B. Derived Verbs

Derived verbs are created by adding morphological form like affixation, reduplication, or compounding to base verbs. These processes alter the meaning or grammatical function of the base verb, resulting in enhanced linguistic variety and contextual precision. Affixation in the *Tulang Bawang* dialect consists of suffixes, prefixes, infixes and confixes. Compound words are the exception in this research. Prefixes are affixes that come at the beginning of words. The types of prefixes are *“di-”, “be-”, “te-”, “pe-”, “pegh-”* while the nasal prefix N- = {*“nge-”, “ng-”, “ny-”, “n-”, “m-”*}.

Table I contains the prefixes and their impact in changing base verb of *Tulang Bawang*. Table II contains the prefixes and their translations in Indonesian. Words end with inflection words end with suffixess. Suffixes include “-ei”, “-ken”, “-nou”, “-meu”, “-keu”, and “-lah”. Table III shows suffixes, and modifying Tulang Bawang's root word has no effect. Infixes go into word roots. Infixes change or create words. The infix is “-em-”. Table IV shows infixes and modifying Tulang Bawang base word has no effect. Additional notes for Table I, where A is an arbitrary letter, C is a consonant letter and V is a vowel letter.

Table V shows three form of reduplication in Tulang Bawang: perfect reduplication, reduplication of the first syllable, and reduplication with affixes. Reduplication of the first syllable occurs if the first syllable has a pattern where the first character is consonant and the second character is 'a', 'u', or 'o', then a new syllable is created with a pattern where the first character is consonant and the second character is 'e' in the initial position of the word. Reduplication with affixes occurs in the first or second word separated by the symbol '-'. Reduplication in Lampung language has a unique pattern.

TABLE I. Prefixes Rules

Rule	*Construct*	*Return*
1	nge{b|d|g|h|l|n|r|w|y}AA...	nge- +{b|d|g|h|l|n|r|w|y}AA...
2	ng{a|i|u|e|o}CV...	ng- +{a|i|u|e|o}CV... | ng- + k{a|i|u|e|o}CV...
3	nyVC...	ny- + cVC... | ny- + sVC...
5	nVC...	n- + tVC...
6	mVC...	m- + pVC...
8	beAA...	be- +AA...
9	peCV...	pe- +CV...
11	perCV...	per- +CV...
12	teAA...	te- +AA..
13	diAA...	di- +AA...

TABLE II. Prefixes in Tulang Bawang

Prefixes	Root Word	Example in Tulang Bawang	Translation in Indonesian
di-	kayun	dikayun	disuruh
be-	tanei	betanei	bertani
te-	alau	tealau	terkejar
pe-	wawai	pewawai	perindah
per-	tego	pertego	pertiga
nge-	nah	ngenah	melihat
ny-	cobou	nyobou	mencoba
ng-	kawil	ngawil	mengail
n-	taban	naban	menggendong
m-	peppul	meppul	membakar

TABLE III. Suffixes in Tulang Bawang

Suffixes	Root Word	Example in Tulang Bawang	Translation in Indonesian
-ei	baccuh	baccuhei	tambahi
-ken	oloh	olohken	kembalikan
-nou	nuwou	nuwounou	rumahnya
-meu	adik	adikmeu	adikmu
-keu	katan	katankeu	lukaku
-lah	mejeng	mejenglah	duduklah

TABLE IV. Infixes in Tulang Bawang

Infix	Root Word	Example in Tulang Bawang	Translation in Indonesian
-em-	cengguk	cemengguk	menunduk

TABLE V. Reduplication in Tulang Bawang

Reduplication	Root Word	Example in Tulang Bawang	Translation in Indonesian
Perfect Reduplication	cobou	cobou-cobou	coba-coba
	nah	nah-nah	lihat-lihat
Reduplication of the initial syllable	dakep	dedakepan	saling berpelukan
	lapah	lelapahan	berjalan-jalan
Reduplication with affixes	balah	bebalah-balah	berkata-kata
	leccak	beleccak-leccakan	berlompat-lompatan

A confix has two parts: one goes on the prefix of a root word and the other on the suffix. Both elements change the base verb or create a new one. Confixes are widespread in Indonesian and regional languages. Confixes in the Tulang Bawang dialect can be seen in Table VI. Words containing a prefix on a confix will be stemmed according to the rules outlined in Table I.

TABLE VI. CONFIXES IN TULANG BAWANG

Confixes	Root Word	Example in Tulang Bawang	Translation in Indonesian
nge-ei	biyak	ngebiyakei	memberati
ng-ei	arit	ngaritei	mengariti
ny-ei	cabut	nyabutei	mencabuti
m-ei	pacul	maculei	mencangkuli
nge-ken	golai	ngegolaiken	menggulaikan
ng-ken	akuk	ngakukken	mengambilkan
ny-ken	sugeu	nyugeuken	menyuguhkan
n-ken	tanem	nanemken	menanamkan
m-ken	pajak	majakken	merebuskan
be-ken	pakkul	bepakkulken	beratapkan
be-an	dakep	bedakepan	berpelukan
di-ken	akuk	diakukken	diambilkan
di-ei	kan	dikanei	dimakani
per-ken	nah	pernahken	perlihatkan
ke-an	kughuk	kekughukan	kemasukan

III. RULE-BASED TULANG BAWANG

The efficacy of stemming a morphology-based word is contingent upon two primary factors: the presence of the base verb in the utilized lexicon and the sequence of the stemming process executed. The sequence of the rule-based stemming process consists of:

1) Execute stemming the reduplication component
2) Execute stemming the confix component
3) Execute stemming the suffix component
4) Execute stemming the prefix component
5) Execute stemming the infix component.

IV. RESEARCH METHODOLOGY

The research stages refer to the steps taken during the research process to obtain the desired results. The steps taken are (1) data collection, (2) word morphology analysis, (3) algorithm design, (4) implementation and testing, and (5) evaluation of results. The principles of the Tulang Bawang dialect stemmer are made with reference to the book *Sistem Morfologi Verba Bahasa Lampung Dialek Tulang bawang*. The research stages is show in Fig. 1.

Detailed explanations related to the TB dialect rule-based stemmer research stages are as follows:

1) *Data collection*: This study employs data collection fundamental word dictionary data, test word data, and independent test word data. Manually entering 6454 terms into the Lampung-Indonesian Dictionary yielded the basic word dictionary information [28]. The data collected for testing the Tulang Bawang Stemmer consists of 500 test words and 200 independent test words according to the morphological rules of Tulang Bawang words.

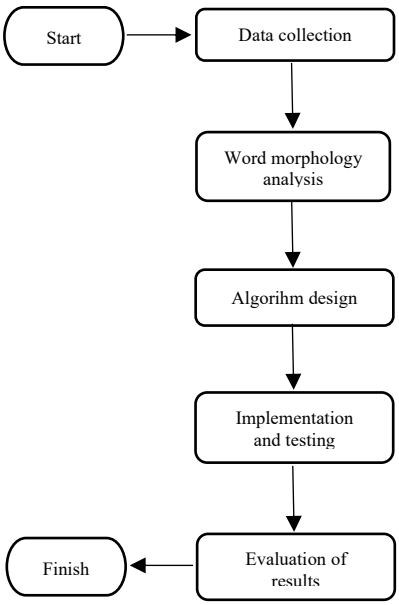

Fig. 1. The research stages

2) *Word morphology analysis*: In the analysis stage, various affixations in the Tulang Bawang dialect are systematically arranged to produce basic words in accordance with the applicable morphological norms/rules. Prefixes, suffixes, infixes, reduplication and confixes used in the Tulang Bawang dialect. The morphological rules obtained include:

a) *Reduplication*: perfect reduplication, reduplication of the initial syllable, reduplication with affixes.

b) *Prefixes*: *di-, be-, te-, pe-, per-, nge-, ny-, ng-, n-, m-*.

c) *Suffixes*: *-ei, -ken, -nou, -meu, -keu, -lah*.

d) *Infixes*: *-em-*.

e) *Confixes*: *nge-ei, ng-ei, ny-ei, m-ei, nge-ken, ng-ken, ny-ken, n-ken, m-ken, be-ken, be-an, di-ken, di-ei, per-ken, ke-an*.

3) *Algorithm design*: The design of the algorithm is the design of the stemming algorithm according to the results of the analysis. The design stage uses pseudocode in detail. The pseudocode for the rule-based TB dialect is provided Fig. 2.

4) *Implementation and Testing*: The application of the algorithm results from the previous step is known as implementation. The pseudocode implementation was written using Python and run on Google Colab.

5) *Evaluation of results*: Stemming evaluation using gold standard assessment is a method for measuring the quality and accuracy of stemming algorithms by comparing the algorithm's stemmed results to a list of accurate base words (gold standard).

6) *Evaluation of results*: Stemming evaluation using gold standard assessment is a method for measuring the quality and accuracy of stemming algorithms by comparing the algorithm's stemmed results to a list of accurate base words (gold standard).

```
Function MAIN_STEMMING(dict_file, in_file, out_file):
   dict = load(dict_file)
   words = load(in_file)
   results = [STEM_WORD(word, dict) for word in words]
   save(results, out_file)
Function STEM_WORD(word, dict):
   // Handle reduplication first
   if is_reduplicated(word):
      return handle_reduplication(word)
  // Check affixes in sequence: confix -> suffix -> prefix -> infix
   patterns = [
      (CONFIX_PATTERNS, remove_confix),
      (SUFFIXES, remove_suffix),
      (PREFIXES, handle_prefix),
      (['em'], remove_infix)
            ]
  for pattern_list, handler in patterns:
      for pattern in pattern_list:
         if matches(word, pattern):
            stemmed = handler(word)
            if stemmed in dict:
               return stemmed
   return word
Function handle_prefix(word):
   prefix = get_prefix(word)
   if prefix == 'nge': return word[3:]
   if prefix == 'ng': return 'k' + word[2:]
   if prefix == 'ny': return ('s'|'c') + word[2:]
   if prefix == 'n': return 't' + word[1:]
   if prefix == 'm': return 'p' + word[1:]
   return word[len(prefix):]
Constants:
   PREFIXES = ['nge','ng','ny','n','m','di','be','te','se','ber','pe','bu']
   SUFFIXES = ['ei','ken','lah','nou','meu','an','keu','no']
```

Fig. 2. The pseudocode rule-based TB dialect

The statistics affixation of the 500 data test are presented in table VII below.

TABLE VII. STATISTICS 500 DATA TEST

Test Word	Total Words
Infixes	9
Confixes	179
Prefixes	174
Reduplication	37
Suffixes	101

The statistics affixation of the 200 independent data test are presented in table VIII below.

TABLE VIII. STATISTICS 200 INDEPENDENT DATA TEST

Test Word	Total Words
Infixes	1
Confixes	36
Prefixes	90
Reduplication	18
Suffixes	55

V. RESULT AND DISCUSSION

The Tulang Bawang Stemmer is developed via a methodology grounded in morphological rules, incorporating language dictionaries. Upon analyzing the current test data, the sequence of deletion employed is reduplication, confixes, suffixes, prefixes, finalizing with the elimination of infixes. The initial step is to ascertain whether the input test word is a root word or not. If the word is absent from the base word dictionary, proceed to remove affixes by verifying the presence of specific affixes in the word. Prior to the removal of an affix, it is verified whether the resultant form is present in the language dictionary. If not, an alternative affix is eliminated. Ultimately, if all affixes are effectively eliminated, the algorithm yields the stem result as a base verb. Alternatively, the algorithm outputs the input word. The Python code was subsequently tested in the Google Colab environment for its implementation. The input consists of 500 test words and 200 independent test words. All evidence from this experiment is archived and accessible at link bit.ly/Zaenal_ICADEIS.

A. Experimental Findings on 500 Test Words

During the implementation stage, the developed Python code was evaluated using 500 samples of test data. The Tulang Bawang dialect is part of Lampung language, which is characterized as a low-resource language so it is not easy to find affixed words on digital media. The test findings showed that out of 500 words, 481 words were accurately stemmed according to the base word reference, while 19 words were not stemmed correctly. The findings are summarized in Table IX. The rule-based methodology achieved 96.2% accuracy due to the careful formulation of rules designed to cover all permutations of affixed words in the Tulang Bawang dialect. To observe the successful stemming of the 481 words based on the reference base verb, the output results are provided at the link bit.ly/Zaenal_ICADEIS.

TABLE IX. STEMMING RESULT IN 500 DATA TEST

Test Word	Rule-based Tulang Bawang Stemmer
Failed Stemming Words	19
Successful Stemming Words	481
Total Words	500
Accuracy	**0.962**

The statistics affixation of the 19 failed words are presented in Table X. Stemming failures have occurred in some words containing prefixes and reduplications.

TABLE X. STATISTICS 19 FAILED WORDS

Test Word	Total Words
Infix	0
Circumfix/Confix	0
Prefix	9
Reduplication	10
Suffix	0

B. Experimental Findings on 200 Independent Test Words

Python code is tested using 200 independent test data during implementation. The two hundred independent test data were different affixation words from the 500 test data and were not easy to obtain. The test results showed that 188 of 200 words were stemmed correctly using the expected base verb reference, while 12 were not.

Table XI showed the outcomes. The rule-based methodology had 94% accuracy due to rigorous rule formulation to encompass all Tulang Bawang dialect affixed word changes. The output showed 188 words stemmed from the reference base verb is available at bit.ly/Zaenal_ICADEIS.

TABLE XI. STEMMING RESULT IN 200 DATA TEST

Independent Test Word	Rule-based Tulang Bawang Stemmer
Failed Stemming Words	12
Successful Stemming Words	188
Total Words	200
Accuracy	**0.94**

The statistics affixation of the 12 failed words are presented in Table XII. Stemming failure had occurred in some words containing all affixations and reduplications.

TABLE XII. STATISTICS 12 FAILED WORDS

Test Word	Total Words
Infix	1
Confix	7
Prefix	2
Reduplication	1
Suffix	2

Tables XIII presented the unsuccessful outcomes of experiment. Despite achieving a high accuracy rate, nearly 100 %, it is crucial to investigate the reasons for the failures in Table XIII. The words '*berpisah*', '*keliwat*', '*perlemou*', and '*pertegeou*' failed to stem because there is no rule for the prefixes '*ber*', '*ke*' and '*per*'. While the terms '*dedakepan*', '*kerau-kerauan*', and '*lelapahan*' failed stemming because the reduplication criteria could not detect modifications in the first syllable. The words '*memejengan*', '*memusikan*', '*ngatei-atei*', '*ngayung-kayung*', '*ngenah-nah*', '*nginum-inum*', and '*ngiset-iset*' are identified as having the prefixes '*m*', '*ng*', or '*ny*' which is erroneous.These words are essentially reduplications, particularly reduplications with alterations to the first syllable and reduplications with prefixes. The words '*ngekui*', '*ngelak*', '*ngenek*' and '*ngepeh*' are identified as having the prefix '*nge-*', which is incorrect; they actually have the prefix '*ng-*'. The order of prefixes in N-prefixes {'*nge-*', '*ng-*', '*ny-*', '*n-*', '*m-*'} should be revisited.

TABLE XIII. STEMMING RESULT THAT FAILED IN 500 DATA TEST

No	Test Word	Stemming Result	Stemming Process
1	berpisah	rpisah	Prefix be removed
2	dedakepan	dedakepan	No change
3	keliwat	keliwat	No change
4	kerau-kerauan	kerau-kerauan	No change
5	lelapahan	lelapahan	No change
6	memejengan	pemejengan	Prefix m removed
7	memusikan	pemusikan	Prefix m removed
8	ngatei-atei	katei-atei	Prefix ng removed
9	ngayung-kayung	kayung-kayung	Prefix ng removed
10	ngekui	kui	Prefix nge removed
11	ngelak	lak	Prefix nge removed
12	ngenah-nah	nah-nah	Prefix nge removed
13	ngenek	nek	Prefix nge removed (result is in the dictionary)
14	ngepeh	peh	Prefix nge removed
15	nginum-inum	kinum-inum	Prefix ng removed
16	ngiset-iset	kiset-iset	Prefix ng removed
17	nyegigil	cegigil	Prefix ny removed
18	perlemou	rlemou	Prefix pe removed
19	pertegeou	rtegeou	Prefix pe removed

Table XIV showed some of the failures that occurred throughout the 200 independent test words. Some of the faults are comparable to those in Table XIII. The words '*cacakanno*' and '*dilemno*' were unable to be stemmed because the model cannot recognize the suffix '*no*'. The words '*melakeuken*', '*ngedidikno*', '*ngehindaghko*', and '*ngejawehko*' were not stemmed due to the model recognizing them as prefixes '*m*' and '*nge*', as well as a lack of rules for the confixes '*me-ken*', '*nge-ko*', and '*nge-no*'. The model failed to recognize the infix -em- in the word '*semuluh*'. The words '*pelulihan*', "*pemasso'an*", and '*pemadem*' were not stemmed since the model identified them as 'pe' prefixes and the rule of '*pe-an*' suffix and '*pem*' prefix was not included in the model. All these facts are very important for further research to improve with other stemming methods.

TABLE XIV. STEMMING RESULT THAT FAILED IN 200 DATA TEST

No	Test Word	Stemming Result	Stemming Process
1	cacakanno	cacakanno	No changes
2	dilemno	lemno	Prefix di removed
3	jejamo	jejamo	No change
4	kelewat	kelewat	No change
5	melakeuken	pelakeuken	Prefix m removed
6	ngedidikno	didikno	Prefix *nge* removed
7	ngehindaghko	hindaghko	Prefix *nge* removed
8	ngejawehko	jawehko	Prefix *nge* removed
9	pelulihan	lulihan	Prefix pe removed
10	pemadem	madem	Prefix pe removed
11	pemasso'an	masso'an	Prefix pe removed
12	semuluh	muluh	Prefix se removed

Some of the limitations of the rule-based approach in the Tulang Bawang dialect based on experiments that have been carried out are (1) the model has not been able to perform stemming on words containing prefixes or suffixes of more than one affix, (2) the lack of availability of basic words or basic verbs that are used as a reference for stemming results, (3) the model has not included the *pe-N* rule pattern, namely {'*penge-*','*peng-*', '*peny-*', '*pen-*', '*pem-*'}, prefix '*per-*', prefix {'*nge-*','*ng-*', '*ny-*', '*n-*', '*m-*'} followed by suffix '*i*', '*ko*', '*no*' and (4) the rules have not worked optimally on reduplication of the initial syllable, reduplication with affixes.

This rule-based stemming of the Tulang Bawang dialect is a preliminary study that was developed using rule information based on word morphology. The original investigation yielded accuracy of 96.2 % on 500 test words and 94 % on 200 independent test words. A initial step towards conducting stemming study using different methodologies. Further stemming research on the Tulang Bawang dialect has the potential to be conducted utilizing the Nazief-Adriani technique, Confix-Stripping, Enhanced Confix-Stripping, N-Gram Stemming, and Syllable pattern in the future.

VI. CONCLUSION

The stemming experiment of affixed words in the Tulang Bawang dialect, utilizing a morphology-based rule-based approach, yielded favorable results, achieving an accuracy of 96.2 % on 500 test words and 94 % on 200 independent test words, as evaluated by the gold standard evaluation method. The findings from this stemming research can be applied in the investigation of Lampung language machine translation, information retrieval, or other NLP case analysis. Future research on the Tulang Bawang dialect is feasible due to the unambiguous identification of some failure factors during stemming.

ACKNOWLEDGMENT

We would like to thank Universitas Teknokrat Indonesia, Universitas Lampung and Kantor Bahasa Lampung for supporting our PhD study in stemming and text data augmentation.

REFERENCES

[1] N. E. Rusminto, I. Rejono, I. Natamenggala, and Sumarti. "Kata tugas bahasa Lampung dialek Tulang Bawang", Jakarta, Indonesia: Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan Nasional, 2000.

[2] W. Hermawan, N. Eko, N. Udin, W. Akhyar, and E. Sanusi. "Sistem Morfologi Verba Bahasa Lampung Dialek Tulang Bawang", Jakarta, Indonesia: Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan Nasional, 2001.

[3] A. Jabbar, S. Iqbal, M. I. Tamimy, A. Rehman, S. A. Bahaj, and T. Saba, "An Analytical Analysis of Text Stemming Methodologies in Information Retrieval and Natural Language Processing Systems," *IEEE Access*, vol. 11, pp. 133681–133702, 2023, doi: 10.1109/ACCESS.2023.3332710.

[4] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[5] A. F. Aji, G. Winata, F. Koto, S. Cahyawijaya, A. Romadhony, R. Mahendra, K. Kurniawan, D. Moeljadi, R. E. Prasojo, T. Baldwin, J. H. Lau, and S. Ruder, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," In *2022 Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May. 2022, pp. 7226-7249. doi : 10.18653/v1/2022.acl-long.500

[6] F. Alam, S. A. Chowdhury, S. Boughorbel, and M. Hasanain, "LLMs for Low Resource Languages in Multilingual, Multimodal and Dialectal Settings," 2024. [Online]. Available: https://aclanthology.org/2024.eacl-tutorials.5/

[7] H. H. Nigatu, A. Lambebo Tonja, B. Rosman, T. Solorio, and M. Choudhury, "The Zeno's Paradox of 'Low-Resource' Languages," 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.983/

[8] A. S. Doğruöz and S. Sitaram, "Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights," 2022. [Online]. Available: https://aclanthology.org/2022.sigul-1.12/

[9] J. Singh and V. Gupta, "A systematic review of text stemming techniques," *Artif Intell Rev*, vol. 48, no. 2, pp. 157–217, Aug. 2017, doi: 10.1007/s10462-016-9498-2.

[10] J. Singh and V. Gupta, "Text stemming: Approaches, applications, and challenges," *ACM Comput Surv*, vol. 49, no. 3, Sep. 2016, doi: 10.1145/2975608.

[11] A. Jabbar, S. Iqbal, M. I. Tamimy, S. Hussain, and A. Akhunzada, "Empirical evaluation and study of text stemming algorithms," *Artif Intell Rev*, vol. 53, no. 8, pp. 5559–5588, Dec. 2020, doi: 10.1007/s10462-020-09828-3.

[12] A. Ardiyanti Suryani, D. Hendratmo Widyantoro, A. Purwarianti, and Y. Sudaryat, "The rule-based sundanese stemmer," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 4, Jul. 2018, doi: 10.1145/3195634.

[13] A. Maesya, Y. Arifin, A. Zahra, and W. Budiharto, "Development of Sundanese Stemmer Based on Morphophonemics," in *10th International Conference on ICT for Smart Society, ICISS 2023 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICISS59129.2023.10291840.

[14] A. Sutedi, R. Elsen, and M. R. Nasrulloh, "Sundanese Stemming using Syllable Pattern," *Jurnal Online Informatika*, vol. 6, no. 2, p. 218, Dec. 2021, doi: 10.15575/join.v6i2.812.

[15] I. Setiawan and H. Y. Kao, "SUSTEM: An Improved Rule-based Sundanese Stemmer," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 6, Jun. 2024, doi: 10.1145/3656342.

[16] N. H. Hrp, M. Fikry, and Y. Yusra, "Algoritma Stemming Teks Bahasa Batak Angkola Berbasis Aturan Tata Bahasa," *Journal of Computer System and Informatics (JoSYC)*, vol. 4, no. 3, pp. 642–648, May 2023, doi: 10.47065/josyc.v4i3.3458.

[17] A. Guterres, Gunawan, and J. Santoso, "Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based," *Teknika*, vol. 8, no. 2, pp. 142–147, Oct. 2019, doi: 10.34148/teknika.v8i2.224.

[18] R. Sovia, S. Defit, and Yuhandri, "Development of the Minangkabau Local Language Translation Machine Based on Stemming," in *Proceeding - 2022 International Symposium on Information Technology and Digital Innovation: Technology Innovation During Pandemic, ISITDI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 195–198. doi: 10.1109/ISITDI55734.2022.9944457.

[19] R. Sovia, S. Defit, Yuhandri, and Sulastri, "Development of natural language processing on morphology-based Minangkabau language stemming algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 1, pp. 542–552, Jul. 2023, doi: 10.11591/ijeecs.v31.i1.pp542-552.

[20] S. H. Wibowo and S. Wibowo, "Development of Stemming Algorithm for Rejang Language Stemmer Based on Rejang Language Morphology View project Development of Stemming Algorithm for Rejang Language Stemmer Based on Rejang Language Morphology," *Article in Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, 2019, [Online]. Available: https://www.researchgate.net/publication/341307354

[21] S. H. Wibowo, R. Toyib, M. Muntahanah, and Y. Darnita, "Time complexity in rejang language stemming," *JURNAL INFOTEL*, vol. 14, no. 3, pp. 174–179, Aug. 2022, doi: 10.20895/infotel.v14i3.764.

[22] F. Amin, W. Hadikurniawati, S. Wibisono, H. Februariyanti, and J. S. Wibowo, "A Hybrid Method of Rule-Based and String Matching Stemmer for Javanese Language," *J Theor Appl Inf Technol*, vol. 15, p. 19, 2017, [Online]. Available: www.jatit.org

[23] P. G. S. C. Nugraha and N. W. Wardani, "Stemming Dokumen Teks Bahasa Bali Dengan Metode Rule Base Approach," 2020. [Online]. Available: http://jurnal.mdp.ac.idjatisi@mdp.ac.idceivedJune1ssedJu

[24] M. Agus, P. Subali, and C. Fatichah, "Kombinasi Metode Rule-Based dan N-Gram Stemming untuk Mengenali Stemmer Bahasa Bali," vol. 6, no. 2, pp. 219–228, 2019, doi: 10.25126/jtiik.201961105.

[25] F. H. Rachman, N. Ifada, S. Wahyuni, G. D. Ramadani, and A. Pawitra, "ModifiedECS (mECS) Algorithm for Madurese-Indonesian Rule-Based Machine Translation," in *2022 International Conference of Science and Information Technology in Smart Administration, ICSINTESA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 51–56. doi: 10.1109/ICSINTESA56431.2022.10041470.

[26] Z. Abidin, A. Wijaya, and D. Pasha, "Aplikasi Stemming Kata Bahasa Lampung Dialek Api Menggunakan Pendekatan Brute-Force dan Pemograman C#," *Jurnal Media Informatika Budidarma*, vol. 5, no. 1, p. 1, Jan. 2021, doi: 10.30865/mib.v5i1.2483.

[27] F. Ariyani, "Distribusi Verba Berprefiks {N-} pada Bahasa Lampung dalam Kitab Kuntara Raja Niti DAN Buku Ajar: Kajian Morfologi," *Jurnal Ranah: Jurnal Kajian Bahasa*, vol. 3, no.2, pp. 124-134, Jul. 2014, doi: 10.26499/rnh.v3i2.43.

[28] Y. Zawarnis, E. M. Kastri, H. Nasution, and A. Saputri. "Kamus Lampung-Indonesia", 2nd ed., Lampung, Indonesia: Kantor Bahasa Provinsi Lampung, 2020.