Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023105028
Vol. 10, No. 1, Februari 2023, hlm. 35-42 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

MODIFIKASI FONEM VOKAL PADA STEMMING KATA TIDAK BAKU

Ahmad Fikri Iskandar[*1], Ema Utami[2], Wahyu Hidayat[3], Agung Budi Prasetio[4], Anggit Dwi Hartanto[5]

[1,2,3,4,5]Univeristas Amikom Yogyakarta, Yogyakarta
Email: [1]andfikri@gmail.com, [2]ema.u@amikom.ac.id, [3]wahyuhublaa@gmail.com,
[4]agung.bp@excelindo.co.id, [5]anggit@amikom.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 21 Mei 2021, diterima untuk diterbitkan: 27 Februari 2023)

Abstrak

Bahasa Indonesia termasuk bahasa yang paling populer digunakan di dunia. Bahasa Indonesia dapat berupa bahasa baku dan tidak baku. Bahasa tidak baku dapat dikarenakan oleh penyerapan dari bahasa asing atau bahasa daerah. Penyerapan ini dapat terjadi perganti huruf vokal. Kontribusi pada penelitian ini adalah melakukan modifikasi fonem pada huruf vokal untuk mengembalikan kata tidak baku ke dalam bentuk kata dasar yang baku disebut sebagai *Modified Vocal Phonemes Non Formal*. Percobaan dilakukan dengan 60 kata tidak baku yang sudah dilakukan *preprocessing* pada penelitian sebelumnya terlebih dahulu. Penelitian ini membandingkan hasil algoritma dengan algoritma pada penelitian sebelumnya. Algoritma *Modified Vocal Phonemes Non Formal* telah berhasil melakukan *stemming* dengan presisi 90.00% dengan 54 kata tidak baku yang sukses dikonversi ke kata dasar sesuai dengan Kamus Besar Bahasa Indonesia (KBBI) dan 6 kata masih belum berhasil dikonversi.

Kata kunci: *fonem vocal, stemming, nazief & adriani, text preprocessing*

MODIFIED VOCAL PHONEMES IN NON-FORMAL WORD STEMMING

Abstract

Indonesian is one of the most popular languages spoken in the world. Indonesian can be standard and non-standard language. Non-standard language can be caused by absorption of foreign languages or village languages. This absorption can occur as a substitute for vowels. The contribution to this research is to modify the phonemes of vowels to return non-formal words into formal root forms known as Modified Vocal Phonemes in Non-Formal. The experiment was carried out with 60 non-formal words that have been preprocessed in the previous research. This research will compare the results of the algorithm with the algorithm in previous research. Algorithm Modified Vocal Phenomes Non-Formal has succeeded in performing stemming with 90.0% precision with 54 words that were successfully converted to base words according to the Big Indonesian Dictionary and 6 words were still not converted.

Keywords: *vocal phonemes, stemming, nazief & adriani, text preprocessing*

1. PENDAHULUAN

Bahasa Indonesia digunakan lebih dari 199 juta orang, sehingga menjadikan Bahasa Indonesia masuk menjadi *Top* 10 bahasa popular digunakan di dunia (Vocket, 2020). Media sosial sekarang ini, menyumbang penggunaan Bahasa Indonesia tidak baku, melalui percakapan langsung (*Direct Message*) dan memberikan postingan atau komen (Putra et al., 2019). Beberapa kata tidak baku biasanya dapat disebut sebagai "*slang*" atau "*alay*".

Kombinasi dari bahasa baku dan tidak baku akan meningkatkan kosa kata pada Bahasa Indonesia. Adapun bahasa tidak baku ini disebabkan dari penyerapan dari bahasa asing (Putradi, 2016) dan bahasa daerah (Prasetyowati, 2020). Bahasa Indonesia digunakan dalam pergaulan di internasional sehingga bahasa asing tidak dapat dihindari. Pengaruh timbal balik dalam komunikasi ini menyebabkan kemajuan pesat dalam bahasa Indonesia, terutama dalam hal peningkatan kosa kata (Putradi, 2016).

Dalam kehidupan sehari-hari, masyarakat Indonesia belum sepenuhnya menggunakan Bahasa Indonesia, karena kosakata bahasa Indonesia itu sendiri dipengaruhi oleh bahasa ibu atau bahasa daerah (Nurhapitudin, 2017). Beberapa perubahan dari bahasa daerah yang digunakan dalam bahasa Indonesia yaitu: 1) pengucapan vokal /u/ → /o/ seperti kata "terus" menjadi "teros" dan "agustus"menjadi "agostus"; 2) pengucapan diftong /ey/ → /ai/ seperti kata "sebagai" menjadi "sebagei"

dan “mulai” menjadi “mulei”; 3) pengucapan diftong /au/ → /o/ seperti kata “saudara” menjadi “sodara” dan “jikakalau” menjadi “jikakalo” (Prasetyowati, 2020).

Struktur afiks Bahasa Indonesia terdiri dari prefiks, infiks, sufiks, dan confix. Afiks adalah morfem yang ditambahkan ke kata untuk membuat kata baru. Awalan adalah imbuhan yang ditambahkan di awal kata. Infiks adalah imbuhan yang disisipkan ke dalam kata dan sufiks adalah imbuhan yang ditambahkan di akhir kata. Sedangkan confix merupakan gabungan dari prefiks dan sufiks dalam sebuah kata (Setiawan et al., 2016).

Analisis teks dengan *Natural Language Processing* (NLP) digunakan untuk kategorisasi teks, peringkasan, analisis sentimen sehingga peneliti harus melakukan normaliasi terhadap kata tersebut. Salah satu pengembangan awal yang sangat populer digunakan terhadap NLP pada *text preprocessing* berbahasa Indonesia adalah *Stemming* Nazief & Adriani (Adriani et al., 2007). Kemudian dikembangkan dengan metode Flexible Affix (Setiawan et al., 2016), tetapi Algoritma *stemmer* ini belum dapat melakukan *stemming* pada kata yang ditempelkan tidak formal, seperti kata “nyimpan” yang memiliki bentuk formal "menyimpan" dan akar kata "simpan".

Permasalahan kata tidak baku sudah diselesaikan oleh Putra & Utami, (2018) dengan menggunakan Algoritma Non-Formal Affix. Algortima ini melakukan aturan dengan menghapus affix pada kata-kata tidak baku, sehingga algoritma ini sukses mengubah kata tidak baku seperti “nyimpan” menjadi kata dasar “simpan”. Walapun begitu, algoritma ini belum sepenuh efektif dalam melakukan memecahakan masalah *non-formal stemming*. Hal didapat dilihat pada kata “*nyimpen*” , algortima tidak dapat melakukan *stemming* dengan baik. Untuk mengatasi hal tersebut Putra et al., (2019) melakukan pendekatan dengan menggunakan *string matching* dengan Levenshtein, tetapi belum terjadi perubahan terhadap hasil kata tersebut sehingga belum mendapatkan performansi terbaiknya. Terakhir Qulub, Utami, and Sunyoto (2020) , menguji dengan algortima Jaro Winkler, tetapi belum mendapatkan hasil perubahan terhadap kata tersebut. Bahkan memberikan hasil kesamaan dengan kata “pimpin”.

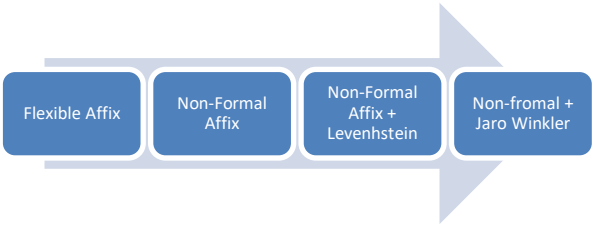
Daftar kata tidak baku yang masih belum berhasil dilakukan *stemming* oleh Putra et al., (2019), yaitu “*Nyebrang*”, “*Critain*”, “*Temenan*”, “*Tukeran*”, “*Mentingin*”, “*Nyelametin*”, “*Nyempetin*”, “*Ngorbanin*”, “*Ngadepin*”, “*Beneran*”, “*Ginian*”, dan “*Cepetan*”. Sebagai contoh kata “*Temenan*” tidak miliki arti pada KBBI. Kata tersebut telah termodifikasi oleh bahasa daerah atau bahasa asing. Kata “*Temenan*” memiliki arti secara formal adalah “Teman”. Hal yang sama dapat dilihat juga pada kata lainnya yaitu “*Beneren*” yang mempunyai arti formal “Benar”.

1.1. Penelitian Terdahulu

Stemming memiliki peran penting dalam melakukan *text processing*. Penelitian *text mining* dengan menggunakan *Stemming* pernah dilakukan untuk *Document Similarity* (Iriananda et al., 2017), *Sentiment Analysis* (Negara et al., 2020), *Personality Detection* (Iskandar et al., 2020) , *Hoax Detection* (Prasetyo et al., 2018) dan lainnya. Beberapa peneliti mencoba untuk meningkatkat algoritma ini dengan beberapa model klasifikasi. *Stemming* memberikan dampak terhadap perfomansi waktu (*cost reduction*) dalam menjalankan model klasifikasi seperti Naïve Bayes (Winarti et al., 2021), K-NN (Utomo et al., 2020), SVM (Utami et al., 2019), dan model lainnya.

Penelitian sebelumnya terkait *stemming* dapat dikelompok menjadi 2 yaitu *Formal Stemming Algorithm* seperti *Flexible Affix Algorithm* (Setiawan et al., 2016) dan *Non-Formal Steemming* seperti *Non-Formal Affix Algortithm* (Putra & Utami, 2018), *Non-Formal Affix* dengan Levenstein (Putra et al., 2019) dan *Non-Formal* dengan Jaro Winkler (Qulub et al., 2020)). Pada *Flexible Affix Algortihm*, prefiks yang digunakan dibagi menjadi panjangnya karakter pada prefiks misal “meng” mempunyai panjang karakter 4 maka maksud kedalam kelompok prefix4. Hal yang sama dilakukan juga untuk sufiks (“kan” akan dikelompokan pada suffix3) (Setiawan et al., 2016). Algoritma ini memulai dengan mengecek kata di kata dasar KBBI, jika tidak terdapat pada kamus KBBI maka akan lanjut ke pengecekan Prefiks dan sufiks. Pengecekan prefiks dilakukan 3 yaitu *delete, addition* dan *check*, sedangkan pengecekan sufiks hanya *delete* dan *check* (Setiawan et al., 2016). Hal ini membuat terdapat kata non-formal yang tidak ternormalisasi pada *Stemming Flexible Algorithm*.

Algoritma *Non-Formal Affix* (Putra & Utami, 2018) merupaka lanjutkan pada algoritma sebelumnya. Algortima melakukan dengan penambahan prefiks, sufiks dan konfiks berdasarkan Zen (2011). Pengujian algoritma ini dilakukan pada 60 kata tidak baku, dengan hasil presisi 73.33% dengan 0 *understemming* 1 *overstemming*, 15 *unstemmed.* Selanjutnya, pendekatan *String Matching* digunaknan pada proses *Stemming* untuk mendapatkan kata-kata yang mungkin *typo* atau serupa dengan menggunakan algoritma Leventein dan Jaro Winkler (Qulub et al., 2020). Pendekatan Levenshtein *distance* yang dilakukan oleh Putra et al. (2019) dengan algortima Non-Formal affix mendapatkan presisi sebesar 80.0% dengan 48 dari 60 kata sukses dilakukan stemming, sedangkan presisi algoritma Non-Formal dengan Jaro Winkler mendapatkan presisi sama yaitu 80.00%, setelah dilakukan perhitungan ulang terhadap hasil kata *stemming*.



Gambar 1. *Road Map* Penelitian

Berdasarkan penelitian terdahulu, Penelitian ini akan berkontribusi pada uji coba *Modified Vocal Phonemes* pada kata tidak baku Putra & Utami, (2018). Sebagai tambahan untuk memperdalam hasil dari uji coba, penelitian ini juga akan melakukan uji coba terhadap data hoaks dari Jala Hoaks dan Jabar Saber Hoaks.

1.2. Huruf Fonem Vokal

Huruf digunakan untuk membentuk kata-kata agar memiliki makna dan dapat mewakili sesuatu yang ingin disampaikan (Rahim, 2020). Huruf vokal adalah huruf fonem yang dihasilkan dengan pita suara yang terbuka. Huruf Vokal biasanya dikatakan huruf hidup. Vokal meliputi: a, i, u, e dan o. Putradi (2016) dan Prasetyowati, (2020) menjelaskan perubahan vokal pada bahasa Indonesia dapat terjadi dari penyerapan dari bahasa asing atau bahasa daerah. Pola penyerapan tersebut dapat berupa satu dan dua vokal. Pola ini sangat sering ditemukan dalam kehidupan sehari-hari, seperti huruf "o" menjadi "u" (Prasetyowati, 2020). Perubahan tersebut bisa disebabkan oleh penyerapan bahasa daerah. Aturan perubahan penyerapan satu dan dua vokal dapat dilihat Tabel 1.

Tabel 1. Pola Penyerapan Satu atau Dua Vokal

Aturan	Contoh	
	Kata Awal	Kata Perubahan
/y/ → /i/	system	sistem
/ea/ → / i/	gear	gir
/e / →/o/	amplitude	amplitudo
/e / →/i/	mate	mati
/e / →/ai/	pake	pakai
/oo/ →/ /u/	idioot	idiot
/oe/ → /u/	moer	mur
/o/ →/ /au/	kalo	kalau
/o/ →/ /u/	teros	terus
/e/ → /a/	benter	bentar
Pelesapan vokal e di akhir kata	turbine	turbin

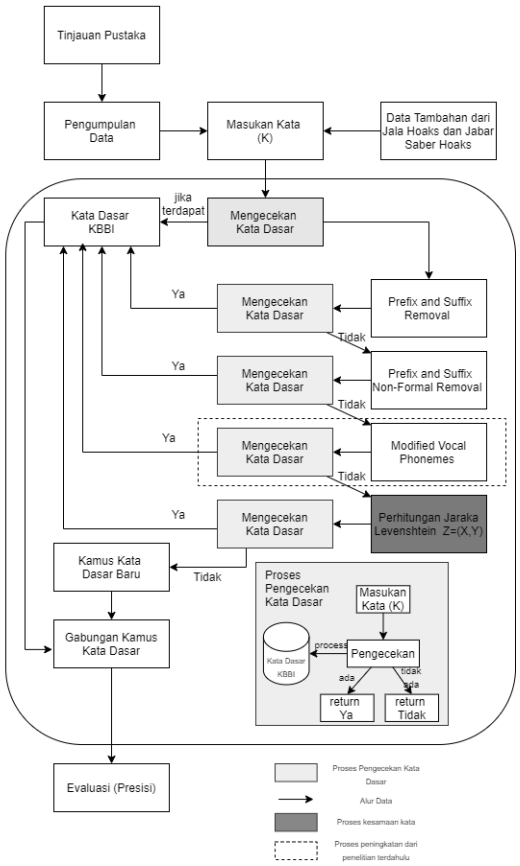
Terdapat 10 aturan pola dasar penyerapan satu atau dua vokal pada **Tabel 1.** Pola penyerapan tersebut akan dilakukan pengujian kata tidak baku.

Berdasarkan permasalahan fonem vokal diatas, penelitian ini melakukan modifikasi fonem pada huruf vokal untuk mengembalikan kata tidak baku ke dalam bentuk kata dasar yang baku berdasarkan KBBI. Kontribusi dari penelitian ini adalah melakukan percobaan modifikasi fonem pada huruf vokal terkait kata tidak baku Bahasa Indonesia atau selanjutnya akan disebut sebagai *Modified Phonemes Vocal.*

Stuktur penulisan pada penelitian ini adalah: Bagian pertama akan dimulai dengan latar belakang serta menjelaskan *novelty* dari penelitian ini. Bagian kedua menjelaskan metode yang diajukan. Bagian ketiga menjelaskan hasil dan serta pembahasan yang didapatkan. Terakhir, penelitain ini akan ditutup pada kesimpulan dan memberikan rekomendasi terhadap penelitian selanjutnya pada Bagian kelima.

2. METODE PENELITIAN

Data yang digunakan adalah data kata non-formal dari penelitian terdahulu (Putra et al., 2019), di mana data sudah berupa kata tidak baku yang sudah siap untuk diuji. Modifikasi terhadap alur penelitian sebelumnya adalah menambahkan uji coba pola perubahan penyerapan fenom satu atau dua vokal dari penelitian Putra et al. (2019). Alur penelitain dapat dilihat pada Gambar 2.



Gambar 2. Alur Penelitian

Kotak yang memiliki garis putus-putus merupakan proses peningkatan dari penelitian sebelumnya. Penelitian ini akan memodifikasi huruf vokal berdasar ketentuan yang sudah ditetapkan. Proses *checking* kata dasar menggunakan KBBI sebagai rujukan pada penelitian ini. Proses pengecekan ini dilakukan 4 kali yaitu proses proses pertama digunakan pada *formal affix*, kedua menggunaakan *non-formal affix*, ketiga

menggunakan modifikasi vokal fenom dan terakhir untuk *string smiliarity* dengan algoritma Levenstein.

Banyaknya variasi aturan fonem vokal mendorong pendekatan pada penelitian ini untuk membuat kategori aturan dasar berdasarkan jumlah huruf vokal yang akan diubah. Aturan perubahan kategori dapat dilihat pada Tabel 2.

Tabel 2. Kategori Aturan *Vocal Phenomes*

Kategori	Aturan
Aturan Fonem 1	/y/ → /i/
Aturan Fonem 1	/e/ → /a/
Aturan Fonem 1	/e/ → /i/
Aturan Fonem 1	/e / →/o/
Aturan Fonem 1	/e / →/ai/
Aturan Fonem 1	/o/ →/ /au/
Aturan Fonem 1	/o/ →/ /u/
Aturan Fonem 2	/oo/ →/ /u/
Aturan Fonem 2	/oe/ → /u/
Aturan Fonem 2	/ea/ → / i/
Aturan Fonem 3	Pelepasan vokal e di akhir kata

Dari 11 aturan pada Tabel 2., akan dilakukan kelompok yaitu kelompok aturan fonem 1 di mana kelompok tersebut memiliki aturan 1 perubahan 1 kata menjadi 1 atau 2 kata, kelompok aturan fenom 2 merupakan perubahan 2 huruf vokal yang bersamaan, dan ketiga merupakan penghapus huruf vokal e di akhir kata. Selain itu, penelitian ini juga akan melakukan uji coba data hoaks dari Jakarta Lawan Hoaks (Jala Hoaks) dan Jabar Saber Hoaks, berjumlah 16.088 kata.

Confusion matrix mengukur presisi algoritma pada data uji. Presisi dapat didefinisikan sebagai alat pengukur yang merupakan ukuran kinerja paling intuitif dan hanya menggunakan rasio pengamatan yang diprediksi dengan benar terhadap pengamatan aktual yang benar (Persamaan 1).

Tabel 3. *Confusion Matrix*

	Actual Positive	***Actual Negative***
Predict Positive	TP	FN
Predict Negative	FP	TN

Sehingga didapatkan (Persamaan 1):

$$Presisi = \frac{TP}{TP+FP} \times 100\% \quad (1)$$

Tabel *confusion matrix* pada Tabel 3, *True Positive* (TP) menjelaskan kata pada data positif yang diprediksi benar, *True Negative* (TN) menjelaskan kata pada data negatif yang diprediksi benar, *False Postive* (FP) menjelaskan kata pada data positif yang diprediksi salah dan terakhir *False Negative* (FN) menjelaskan kata pada data negatif yang diprediksi salah.

3. HASIL PENELITIAN

4.1. Hasil Stemming

Hasil *stemming* kata tidak baku didapatkan dengan implementasi Algoritma *Modified Vocal Phenomes* dapat dilihat pada Tabel 4.

Tabel 4. *Hasil Stemming Modified Vocal Phenomes*

Kata Dasar	Kata Tidak Baku	Non-Formal Affix + Levenshtein	Modified Vocal Phenomes
terjang	Nerjang	Terjang	**Terjang**
tuduh	Nuduh	Tuduh	**Tuduh**
terima	Nerima	Terima	**Terima**
tegur	negur	Tegur	**Tegur**
pukul	mukul	Pukul	**Pukul**
pimpin	Mimpin	Pimpin	**Pimpin**
coba	nyoba	Coba	**Coba**
siram	nyiram	Siram	**Siram**
suruh	Nyuruh	Suruh	**Suruh**
simpan	Nyimpen	Simpan	**Simpan**
sebrang	Nyebrang	Bang	Nyebrang
anggap	Nganggep	Anggap	**Anggap**
amuk	Ngamuk	Amuk	**Amuk**
ambil	Ngambil	Ambil	**Ambil**
buka	Ngebuka	Buka	**Buka**
bantu	Ngebantu	Bantu	**Bantu**
lepas	Ngelepas	Lepas	**Lepas**
bayang	Kebayang	Bayang	**Bayang**
injak	Keinjek	Injak	**Injak**
kabul	Kekabul	Kabul	**Kabul**
pergok	Kepergok	Pergok	**Pergok**
tipu	Ketipu	Tipu	**Tipu**
ulang	Keulang	Ulang	**Ulang**
wujud	Kewujud	Wujud	**Wujud**
Cerita	Critain	Urita	Critain
betul	Betulin	Betul	**Betul**
manja	Manjain	Manja	**Manja**
ganggu	Gangguin	Ganggu	**Ganggu**
ganti	Gantian	Ganti	**Ganti**
ikut	Ikutan	Ikut	**Ikut**
musuh	Musuhan	Musuh	**Musuh**
sabun	Sabunan	Sabun	**Sabun**
teman	Temenan	Semen	**Teman****
tukar	Tukeran	Teker	**Tukar****
tanya	nanyain	Tanya	**Tanya**
tunjuk	nunjukin	Tunjuk	**Tunjuk**
penting	mentingin	Tingi	Lenting
pegang	megangin	Pegang	**Pegang**
selamat	nyelametin	Gamet	**Selamat****
sempat	nyempetin	Empat	Empet
korban	ngorbanin	Sorban	**Korban****
hadap	ngadepin	Idep	Adap
bukti	ngebuktiin	Bukti	**Bukti**
warna	ngewarnain	Warna	**Warna**
dengar	Kedengeran	Dengar	**Dengar**
temu	ketemuan	Temu	**Temu**
benar	beneran	Tener	**Benar****
begini	ginian	mi	Ginian
kawin	kawinan	Kawin	**Kawin**
main	mainan	Main	**Main**
parkir	parkiran	Parkir	**Parkir**
dulu	duluan	Dulu	**Dulu**
gendut	gendutan	Gendut	**Gendut**
karat	karatan	Karat	**Karat**
paling	palingan	Paling	**Paling**
sabar	sabaran	Sabar	**Sabar**
bagus	kebagusan	Bagus	**Bagus**
sana	sanaan	Sana	**Sana**
cepat	cepetan	repet	**Cepat****
pagi	sepagian	Pagi	**Pagi**

Evaluasi dilakukan dengan menghitung presisi. Kata tebal diatas merupakan hasil *non-formal stemming* yang benar dengan kata dasar pada penelitain sebelumnya (Putra et al., 2019). Kata

dengan tanda dua bintang (**) merupakan kata yang berhasil dilakukan *stemming* pada penelitian ini, sedangan kata tanpa tebal merupakan yang belum berhasil dikonversi ke kata dasar berdasarkan KBBI.

Algoritma Non-formal Affix dengan Levenhstein (Putra et al., 2019) berhasil mendapatkan 48 kata atau 80.00% Terdapat 12 kesalahan pada proses *stemming* (0 *understeeming*, 1 *overstemming*, dan 11 *unstemming*). *Overstemming* kata tidak baku "Ginian" yang seharunya "Begini" menjadi "Mi". sedangkan untuk 11 kata *unstemming* yaitu "*Nyebrang*", "*Critain*", "*Temenan*", "*Tukaran*", "*Mentingin*", *Nyelametin*", "*Nyempetin*", "*Ngorbanin*", "*Ngadepin*", "*Beneran*", dan "*Cepetan*" belum dilakukan perubahan.

Algoritma *Modified Vocal Phenomes* telah berhasil melakukan presisi 90.00% dengan 54 kata berhasil dilakukan *stemming* pada Tabel 4. Kata yang berhasil dilakukan stemming adalah "*Temenan*" menjadi "Teman", kata "*Tukeran*" menjadi "Tukar", "Nyelamtin menjadi "Selamat", "Ngorbanin" menjadi "Korban" dan "Beneran" menjadi "Benar". Sedangkan jumlah kata baku yang berhasil dilakukan *stemming* sebesar 6 (1 *overstemming*, 3 *unstemming*, 2 *understemming*). Kata-kata yang tidak berhasil dilakukan *stemming* adalah "*Nyebrang*", "*Critain*", "*Mentingin*", "*nyempetin*", "*ngadepin*", dan "*Ginian*". Algoritma ini meningkatkan penelitian sebelumnya, yang hanya mendapatkan presisi terhadap kata yang berhasil *stemming* dari 48 pada Putra et al. (2019) menjadi 54 dari 60 tidak baku atau meningkatkan presisi sebesar 12,5%.

4.2. Diskusi

Daftar kata-kata yang sudah berhasil stemming menjadi kata dasar berdasarkan KBBI dengan *Modified Vocal Phenomes* pada Tabel 5.

Tabel 5. *Perbandingan Hasil Stemming Kata Tidak Baku*

Kata Tidak Baku	Hasil Stemming (Putra et al., 2019)	Hasil Stemming (Penelitian ini)
Temenan	Semen	**Teman****
Tukeran	Teker	**Tukar****
nyelametin	Gamet	**Selamat****
ngorbanin	Sorban	**Korban****
beneran	Tener	**Benar****
cepetan	repet	**Cepat****

Faktor yang mempengaruhi keberhasilan *stemming* pada penelitian ini dengan penelitian yang dilakukan oleh (Putra et al., 2019) adalah perubahan huruf vokal pada setiap kata. Algoritma *Modified Phenomes Vocal* sangat efektif jika terdapat penggunaan huruf vokal yang salah. Sebagai contoh kata tidak baku "*Temenan*". Jika mengikuti algoritma Non-Formal Affix (Putra et al., 2019), maka hal yang pertama dilakukan adalah dihapus afiksnya, sehingga menjadi kata "Temen". Selanjutnya dilakukan dengan *word similarity*, maka kata "*Temen*" memiliki kecocokan dengan kata "Semen".

Sedangkan pada penelitian ini, sebelum dilakukan *word similarity* kata "Temen" akan dilakukan modifikasi fonem huruf vokal pada karakter urutan kedua dan keempat. Hasil modifikasi fenom vokal dapat dilihat pada Tabel 6.

Tabel 6. *Modifikasi Fenom Vokal "Temen"*

Modifikasi Fonem Vokal	Pengecekan KBBI	Levenstein "Temenan"
~~Temen~~	~~Tidak ada~~	-
Teman	Ada	2
Temin	Ada	3
~~Temon~~	~~Tidak ada~~	-
~~Temain~~	~~Tidak ada~~	-
~~Tamen~~	~~Tidak ada~~	-
Taman	Ada	3
~~Tamin~~	~~Tidak ada~~	-
~~Tamon~~	~~Tidak ada~~	-
~~Tamain~~	~~Tidak ada~~	-
~~Timen~~	~~Tidak ada~~	-
~~Timan~~	~~Tidak ada~~	-
~~Timin~~	~~Tidak ada~~	-
~~Timon~~	~~Tidak ada~~	-
~~Timain~~	~~Tidak ada~~	-
~~Tomen~~	~~Tidak ada~~	-
Toman	Ada	3
~~Tomin~~	~~Tidak ada~~	-
Tomon	Ada	4
~~Tomain~~	~~Tidak ada~~	-
~~Taimen~~	~~Tidak ada~~	-
~~Taiman~~	~~Tidak ada~~	-
~~Taimin~~	~~Tidak ada~~	-
~~Taimon~~	~~Tidak ada~~	-
~~Taimain~~	~~Tidak ada~~	-

Pada Tabel 6., terdapat 25 kombinasi kata dari hasil modifikasi fenom vokal kata "Temen". Coretan pada kata (*Strikethrough word*) menjelaskan bahwa kata tersebut tidak ada di KBBI. Hanya kata yang ada pada KBBI akan dilakukan *word similarity* dengan Levenstein. Dari kata-kata yang ada di KBBI terdapat kata "Teman" memiliki jarak 2, "Temin" memiliki jarak 3, "Taman" memiliki jarak 3, "Toman" memiliki jarak 3 dan "Tomon" memiliki jarak 4. Pada kasus ini didapatkan kata "Temen" mendekati kata "Teman".

Perbandingan hasil kata tidak baku yang tidak berhasil *stemming* dari penelitian ini dengan penelitian sebelumnya (Putra et al., 2019) dapat dilihat pada Tabel 7.

Tabel 7. *Kata yang tidak berhasil stemming*

Kata Tidak Baku	Hasil Stemming (Putra et al., 2019)	Hasil Stemming (Penelitian ini)
Nyebrang	Bang	Nyebrang
Critain	Urita	Critain
mentingin	Tingi	Lenting
nyempetin	Empat	Empet
ngadepin	Idep	Adap
ginian	mi	Ginian

Selanjutnya kata yang tidak berhasil di *stemming* sebanyak 6 dari 60 kata tidak baku atau error 10%. Dari kata-kata tersebut yang dapat dilihat pada Tabel 7. dapat dikategorikan kedalam *unstemming* ("*Critain*", "*Ginian*", "*Nyebrang*"); *overstemming* (*Lenting*) dan *understemming* ("*empet*", "*adap*"). Faktor yang mempengaruhi

kurangnya hasil *stemming* pada *Modified Vocal Phenomes* adalah kekurangan beberapa karakter pada kata tidak baku dan *competitive similarity distance* yang tinggi.

Faktor kehilangan beberapa karakter baik di awal, tengah atau akhir kata tidak baku mempengaruhi proses stemming, sehingga menghasilkan *unstemming* dan *understemming*. Hasil yang tidak dilakukan perubahan (*unstemming*) seperti, "*Nyebrang*", "*Critain*" kelihangan karakter "e" pada ditengah kata dan "*Gini*" kelihangan dua karakter yaitu "b" dan "e".

Tabel 8. Mising Character

Kata Tidak Baku	Penjelasan	Kata
Nyebrang	"e" pada karakter kelima	Seberang
Critain	"e" pada karakter kedua	Ceritain
Gini	"b" dan "e" pada karakter pertama dan kedua	Begini

Faktor kedua terjadi karena kata tidak baku memilliki *competitive similarity distance*, yaitu kata-kata yang memiliki jarak satu dengan kata yang lain sangat dekat (dalam arti perbedaan hanya satu karakter, bahkan penyebutan kata tersebut hampir sama), seperti kata tidak baku "*Mentengin*" memiliki kata dasar "Penting". Kata "Penting" itu sendiri memiliki jarak yang dekat dengan kata "Denting", "Lenting", "Penting" , "Senting", "Genting" atau "Sinting".

Tabel 9. Kemiripan Kata "Penting"

Kata Baku	Penjelasan
Denting	Perbedaan pada karakter pertama yaitu "D"
Lenting	Perbedaan pada karakter pertama yaitu "L"
Senting	Perbedaan pada karakter pertama yaitu "S"
Genting	Perbedaan pada karakter pertama yaitu "G"
Sinting	Perbedaan pada karakter pertama yaitu "S" dan pada karakter kedua yaitu "I"

Perbedaan diatas, didominasi oleh pada order 1 yaitu pada "D", "L", "S", dan "G" dan pada order 2 ada kata "I". Hal ini sesuai dengan permasalahan kata tidak baku menggunakan *similarity distance* masih terdapat jarak yang cukup signifikatn dekat dengan kata dasar pada KBBI (Putra et al., 2019).

Selanjutnya, pengujian dilakukan pada *text preprocessing* dengan data yang cukup besar *Hoax Detection* dari Jala Hoaks dan Jabar Saber Hoaks. Dari hasil tokenisasi didapatkan unik kata sebanyak 16.088 yang terdiri dari 4055 merupakan kata dasar yang terdapat di KBBI sedangkan 12.033 tidak terdapat di KBBI. Setelah dilakukan *stemming* didapatkan 5490 kata berhasil dilakukan perubahan menjadi kata dasar yang terdapat pada KBBI, dan 6543 masih belum terdaftar pada kata dasar KBBI. Kata yang gagal dilakukan *stemming* merupakan singkatan ("polri", "psbb", "rusd", "polres", "Kominfo", "Kemenkes"), Bahasa Asing ("website", dan "news") dan nama dapat berupa nama perusahaan atau virus ("Google", "Youtube", "Corona", "Instagram") dan lainya. Sehingga perlu dilakukan seleksi kata terlebih dahulu terkait *acronim*, *name of entity* dan bahasa asing.

4. KESIMPULAN

Percobaan *stemming* kata tidak baku menggunakan *Modified Vocal Phenomes* telah berhasil meningkatkan presisi dari penelitian sebelumnya. Hasil implementasi *Modified Vocal Phenomes* masih belum sempurna, dikarena masih terdapat beberapa kata yang "*Nyebrang*", "*Critain*", "*mentingin*", "*nyempetin*", "*ngadepin*", "*Ketemuan*", "*ginian*" dan "*duluan*". Algoritma ini telah berhasil mendapatkan presisi 90.00% dengan 54 kata sukses dilakuakn proses *stemming* dan 6 kata masih belum berhasil dikonversi ke kata dasar

Algoritma ini dipengaruhi oleh 2 faktor yaitu kekurangan beberapa karakter pada kata tidak baku dan *competitive similarity distance* yang tinggi pada kata tidak baku tersebut, sehingga belum bisa melakukan *stemming* pada 6 kata tidak baku yaitu "*Nyebrang*", "*Critain*", "*Mentingin*", "*nyempetin*", "*ngadepin*", dan "*Ginian*". Diharapkan penelitian selanjutnya dapat melakukan modifikasi *additional or insert* karakter pada setiap kata untuk mendapatkan hasil *stemming* yang lebih baik lagi.

DAFTAR PUSTAKA

ADRIANI, M., ASIAN, J., NAZIEF, B., TAHAGHOGHI, S. M., & WILLIAMS, H. E. 2007. Stemming Indonesian: A confix-stripping approach. ACM Transactions on Asian Language Information Processing (TALIP), 6(4), 1-33.

HEINZ, H. C. M. G. H. R. 2013. *Persepsi Masyarakat Terhadap Perawatan Ortodontik Yang Dilakukan Oleh Pihak Non Profesional*, *53*(9), 1689–1699.

IRIANANDA, S. W., MUSLIM, M. A., & DACHLAN, H. S. 2017. *Similarity Based on Class-Based Indexing*. *9*(1).

ISKANDAR, A. F., UTAMI, E., & PRASETIO, A. B. 2020. *Word Analysis of Indonesian Keirsey Temperament*. *14*(4), 365–376.

NEGARA, A. B. P., MUHARDI, H., & PUTRI, I. M. 2020. Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(3), 599. https://doi.org/10.25126/jtiik.2020711947

NURHAPITUDIN, I. 2017. Penggunaan Kosa Kata Bahasa Daerah Dalam Komunikasi Berbahasa Indonesia Sebagai Bahasa Tuturan. *Jurnal Al-Tsaqafa*, *14*, 265–274.

PRASETYO, A. R., INDRIANTI, & ADIKARA, P. P. 2018. Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan

Menggunakan Metode Modified K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(12), 7466–7473.

PRASETYOWATI, R. 2020. *Kesalahan Pengucapan Diftong dan Vokal U pada Pidato Gubernur Jawa Tengah Ganjar Pranowo Dalam Rangka HUT Ke-74 Republik Indonesia*. https://doi.org/10.31219/osf.io/g3f68

PUTRA, R. B. S., & UTAMI, E. 2018. Non-formal affixed word stemming in Indonesian language. *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, *2018-Janua*, 531–536. https://doi.org/10.1109/ICOIACT.2018.8350735

PUTRA, R. B. S., UTAMI, E., & RAHARJO, S. 2019. Accuracy measurement on Indonesian non-formal affixed word stemming with levenhstein. *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 486–490. https://doi.org/10.1109/ICOIACT46704.2019.8938423

PUTRADI, A. W. A. 2016. Pola-Pola Perubahan Fonem Vokal Dan Konsonan Dalam Penyerapan Kata-Kata Bahasa Asing Ke Dalam Bahasa Indonesia: Kajian Fonologi. *Jurnal Arbitrer*, *3*(2), 95. https://doi.org/10.25077/ar.3.2.95-112.2016

QULUB, M., UTAMI, E., & SUNYOTO, A. 2020. Stemming Kata Berimbuhan Tidak Baku Bahasa Indonesia Menggunakan Algoritma Jaro-Winkler Distance. *Creative Information Technology Journal*, *5*(4), 254. https://doi.org/10.24076/citec.2018v5i4.218

SETIAWAN, R., KURNIAWAN, A., BUDIHARTO, W., KARTOWISASTRO, I. H., & PRABOWO, H. 2016. Flexible affix classification for stemming Indonesian Language. *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2016*. doi.org/10.1109/ECTICon.2016.7561257

UTAMI, E., HARTANTO, A. D., ADI, S., OYONG, I., & RAHARJO, S. 2019. Profiling analysis of DISC personality traits based on Twitter posts in Bahasa Indonesia. *Journal of King Saud University - Computer and Information Sciences*, doi.org/10.1016/j.jksuci.2019.10.008

UTOMO, F. S., SURYANA, N., & AZMI, M. S. 2020. Stemming impact analysis on Indonesian Quran translation and their exegesis classification for ontology instances. *IIUM Engineering Journal*, *21*(1), 33–50. https://doi.org/10.31436/iiumej.v21i1.1170

VOCKET. 2020. 10 bahasa yang paling banyak digunakan di Internet. *TheVocket.Com*, 2020. https://www.thevocket.com/10-bahasa-paling-banyak-internet/, akses 6 Juli 2021 jam 10.22.

WINARTI, T., INDRIYAWATI, H., VYDIA, V., & CHRISTANTO, F. W. 2021. Performance comparison between naive bayes and k-nearest neighbor algorithm for the classification of indonesian language articles. *IAES International Journal of Artificial Intelligence*, *10*(2), 452–457. doi.org/10.11591/IJAI.V10.I2.PP452-457

ZEN, E. L. 2011. Afiks Tidak Baku Dalam Bahasa Indonesia Ragam Informal. *LiNGUA: Jurnal Ilmu Bahasa Dan Sastra*, *6*(1). https://doi.org/10.18860/ling.v6i1.1300

Halaman ini sengaja dikosongkan