ISSN: 2477-5126
e-ISSN: 2548-9356

Optimalisasi *Stemming* Kata Berimbuhan Tidak Baku Pada Bahasa Indonesia Dengan *Levenshtein Distance*

Rahardyan Bisma Setya Putra[1*)], Ema Utami[2], Suwanto Raharjo[3]
[1,2]Magister Teknik Informatika, Universitas Amikom Yogyakarta
[3]Teknik Informatika, Fakultas Teknologi Industri, Institut Sains & Teknologi AKPRIND Yogjakarta
[1,2]Condong Catur, Depok, Sleman, Yogyakarta 55281, Indonesia
[3]Jln. Kalisahak No.28 Kompleks Balapan Tromol Pos 45 Yogjakarta, Indonesia
email: [1]amsibsam@mail.com, [2]ema.u@amikom.ac.id, [3]wa2n@akprind.ac.id

Received: 30 Maret 2018; Revised: 10 Mei 2018; Accepted: 14 Mei 2018

***Abstract* – Stemming algorithm Nazief & Andriani has been development in terms of the speed and the accuracy. One of its development is Non-formal Affix Algorithm. Non-formal Affix Algorithm improves the accuracy for non-formal affixed word. In its growth, Indonesian language is used in two ways: formal and non-formal. Non-formal language is commonly used in casual situations such as conversations and social media post (Facebook, Twitter, Instagram, etc.). To get the root of the word of a casual conversation or a social media post, stemming algorithm which can process the non-formal words with affixes already proposed. But, the previous algorithm unable to stem a non-formal word that slightly change the root word. Therefore, this study modifies Non-formal Affix Algorithm to increase stemming accuracy on non-formal word. Modifications are made by adding Levenshtein Distance. The result of the research shows that the algorithm made in this research has 96.6% accuracy while the Non-formal Affix algorithm has 73.3% accuracy in processing 60 non-formal affixed words. Based on the result, Levenshtein Distance approach can increase the accuracy on stemming non-formal affixed word.**

***Abstrak* – Algoritma stemming Nazief & Andriani sudah banyak dikembangkan dari sisi kecepatan maupun akurasi. Salah satu pengembangannya adalah Algoritma Non-formal Affix. Algoritma ini meningkatkan akurasi *stemming* pada kata berimbuhan tidak baku. Pada perkembangannya, Bahasa Indonesia digunakan dalam dua cara: baku dan tidak baku. Bahasa tidak baku biasanya digunakan pada situasi yang santai seperti percakapan santai ataupun post dan komen di sosial media (Facebook, Twitter, Instagram, dll). Untuk mendapatkan kata dasar dari kata berimbuhan tidak baku telah diusulkan sebelumnya algoritma *stemming Non-formal Affix*. Namun algoritma ini masih memilki keterbatasan dalam melakukan *stemming* kata berimbuhan tidak baku yang memiliki sedikit perubahan pada kata dasarnya. Oleh karena itu penelitian ini berfokus pada modifikasi algoritma *Non-formal Affix* untuk meningkatkan akurasinya dalam *stemming* kata berimbuhan tidak baku. Hasil dari penelitian ini menunjukkan bahwa hasil modifikasi dengan *Levenshtein Distance* memiliki tingkat akurasi 96.6%, sedangkan algoritma *Non-formal Affix* memiliki akurasi 73.3% pada saat *stemming* 60 kata berimbuhan tidak baku. Sehingga dapat disimpulkan bahwa pendekatan dengan *Lenshtein Distance* dapat meningkatkan akurasi algoritma *Non-formal Affix* dalam melakukan stemming kata berimbuhan tidak baku.**

***Kata Kunci* – *Natural Language Processing, Stemming, Levenshtein Distance, Similarity*.**

*) **Corresponding author:** Rahardyan Bisma Setya Putra
Email: amsibsam@gmail.com

I. PENDAHULUAN

Bahasa Indonesia dapat digunakan secara formal maupun non-formal. Penggunaan Bahasa non-formal di Indonesia biasanya dilakukan pada kondisi yang santai misalnya seperti pada saat *chat*, komentar di media sosial, atau *post* di media sosial seperti. Orang Indonesia cenderung menggunakan Bahasa non-formal seperti pada cuitan yang diambil dari twitter. Salah satunya adalah "Lu ga bakal bias nemuin rekamannya, ini khusus" kata "nemuin" dimana kata formalnya adalah "menemukan" dengan kata dasar "temu". Twitter dapat menyediakan data dengan jumlah yang besar dan mudah didapat [15]. Sehingga akan banyak dilakukan pengolahan data yang bersumber dari twitter yang dimana kebanyakan masyarakat Indonesia menggunakan Bahasa tidak baku.

Pada penelitian sebelumnya dengan judul "*Non-formal Affixed Word Stemming In Indonesian Language*" menghasilkan modifikasi algoritma *stemming* Nazief & Andriani dengan improvisasi *Flexible Affix Classification* dari Reina Setiawan [1]. Penelitian sebelumnya ini berfokus pada modifikasi algoritma *stemming* Bahasa Indonesia agar dapat melakukan *stemming* pada kata berimbuhan tidak baku [2].

Stemming dapat diterapkan sebagai *text processing*, seperti *information retrieval*, pengecekan *plagiarism*, peningkatan performa pencarian, dan lainnya [3,4,5,6]. *Stemming* pada kata tidak baku dapat dimanfaatkan sebagai bagian dari proses *sentiment analysis, chat bot*, dan pemrosesan teks lainnya [2]. Pada penelitian sebelumnya *stemming* Non-formal Affix masih memiliki keterbatasan. Algoritma Non-formal Affix tidak dapat melakukan stemming pada kata berimbuhan tidak baku yang kata dasarnya sedikit berubah seperti "Critain" dengan kata dasarnya adalah "Cerita".

Berdasarkan pada masalah yang ada pada penelitian sebelumnya, maka penelitian ini akan melakukan modifikasi pada algoritma *Non-formal Affix* agar dapat melakukan stemming pada kata berimbuhan tidak baku yang kata dasarnya sedikit berubah. Dengan saran dari penelitian

sebelumnya, maka pendekatan menggunakan algoritma *edit distance (Levenshtein)* untuk menghitung kemiripan kata dasar yang ada pada *dictionary* dengan kata dasar yang sedikit berubah. *Edit distance* dipropos oleh ilmuan Rusia Vladimir Levenshtein pada tahun 1965 [7]. Algoritma *Levenshtein* adalah pengukur kemiripan antara dua *string* [8]. Diharapkan modifikasi ini dapat meningkatkan akurasi dari algoritma *Non-formal Affix* dalam melakukan stemming kata berimbuhan tidak baku.

II. PENELITIAN YANG TERKAIT

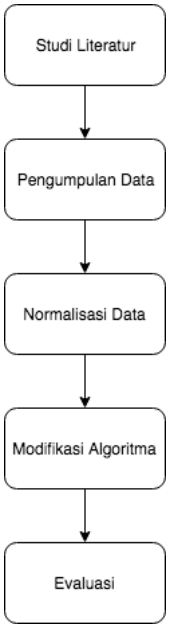
Banyak penelitian yang mengembangkan atau menggunakan stemming. Salah satu algoritma stemming pada Bahasa Indonesia yang banyak dikembangkan adalah algoritma Nazief & Andriani. *Flexible Affix Classification* meningkatkan performa algoritma Nazief & Andriani dalam melakukan stemming kata yang berulang seperti "berlari-lari", "bersalam-salaman". Dari 1704 text pada dokumen yang di stemming menunjukkan bahwa algoritma ini memiliki hasil akurasi yang lebih baik. Penelitian tersebut dilakukan oleh Setiawan, Kurniawan, Budiharto, Kartowisastro, dan prabowo [1]. Algoritma Nazief & Andriani juga digunakan sebagai *retrieval system* dalam Al Hadth pada Bahasa Indonesia oleh Atqia Aulia, Dewi Khairani, dan Nashrul Hakiem [9]. Pada penelitian ini dilakukan penggabungan teknik *stemming* Nazief & Andriani dengan Bahasa pemrogramman PHP untuk menampilkan hasil pencarian hadis. Penelitian tersebut menghasilkan 1 recall score dan precision 96.1%.

Asian, William, dan Tahaghohhi mengembangkan teknik *confix-stripping* pada algoritma Nazief & Andriani [10]. Penelitian tersebut menghasilkan peningkatan akurasi dari Nazief & Andriani 93% menjadi 95%. Confix-stripping juga dilakukan oleh Arifin, Mahendra, dan CIptaningtyas [11]. Penelitian tersebut sukses meningkatkan pengurangan term size menjadi 32.66% dari sebelumnya adalah 30.95%. Rahardyan dan Ema Utami mengembangkan algoritma Nazief & Andriani dengan menambahan Non-formal Affix rule agar algoritma Nazief & Andrian dapat melakukan stemming pada kata berimbuhan tidak baku [2]. Penelitian tersebut berhasil meningkatkan akurasi dari algoritma Nazief & Andriani dalam melakukan stemming kata tidak baku dari 35% menjadi 73.3%. Namun pada penelitian tersebut masih terdapat beberapa kegagalan hasil *stemming* yang dikarenakan kata berimbuhan tidak baku yang sedikit merubah kata dasarnya. Oleh karena itu penelitian yang akan dilakukan saat ini adalah memodifikasi algoritma non-formal affix dengan pendekatan *Levenshtein Distance* agar dapat meningkatkan akurasi dalam melakukan stemming kata berimbuhan tidak baku

III. METODE PENELITIAN

A. Tahapan Penelitian

Penelitian yang dilakukan berfokus pada modifikasi algoritma *Non-formal Affix* dengan melakukan penambahan pendekatan string similarity menggunakan algoritma *Levenshtein Distance*. Alur penelitian dapat dilihat pada Gbr. 1.

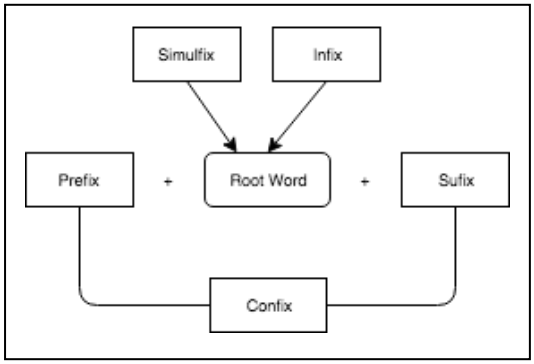


Gbr. 1 Alur penelitian

1) Melakukan studi literatur untuk memperdalam pengetahuan dan mencari landasan teori mengenai struktur imbuhan pada Bahasa Indonesia, imbuhan tidak baku, *stemming*, dan *levenshtein distance*.
2) Data merupakan jenis sekunder yang diambil dari penelitian sebelumnya yang berjudul "Non-formal Affixed Word Stemming in Indonesian Language" [2] dari sumber pertama dari penelitian "Afiks Tidak Baku dalam Bahasa Indonesia Ragam Informal" [14].
3) Data dinormalisasi ke dalam bentuk yang kompatibel untuk dilakukan proses dalam Bahasa pemrogramman.
4) Melakukan modifikasi algoritma pada penelitian "Non-formal Affixed Word Stemming in Indonesian Language" dengan melakukan penambahan pendekatan *Levenshtein Distance*. Cara kerja dapat dilihat pada Gbr 4.
5) Dilakukan evaluasi untuk mengukur tingkat akurasi dari hasil modifikasi algoritma yang ditambahkan pendekatan *Levenshtein Distance*. Evaluasi dilakukan dengan cara membandingkan persentase keberhasilan antara algoritma *Non-formal Affix* dengan algoritma yang sudah dimodifikasi dengan *Levenshtein Distance*.

B. Struktur Imbuhan Bahasa Indonesia

Pada penelitian sebelumnya dikembangkan algoritma dengan algoritma dasar yaitu Nazief & Andriani. Pengembangan yang dilakukan menggunakan meningkatkan akurasi algoritma Nazief & Andriani untuk melakukan *stemming* pada kata berimbuhan tidak baku. Berdasar penelitian sebelumnya, struktur imbuhan bahasa Indonesia terdiri dari prefiks, infiks, sufiks, konfiks, dan simulfiks [2]. Struktur imbuhan pada Bahasa Indonesia dapat dilihat pada Gbr. 2.



Gbr. 2 Struktur imbuhan Bahasa Indonesia

1) Prefiks: Prefiks adalah afiks yang diletakkan di depan sebuah kata dasar. Menurut S. Takdir Alisjahbana di dalam Khotimah, prefiks *di-, ke-, ter-*. Memiliki kegunaan untuk menyatakan tempat dan bentuk positif [12]. Example: *me- (menjual), ke- (kedepan), ter- (termakan), per- (perkuda), se- (sebagai), ber- (berjalan).*

2) Sufiks: Sufiks merupakan afiks yang diletakkan di bagian belakang kata dasar [12]. E.g.: *-an (makanan), -i (manusiawi), wi* merupakan perubahan bunyi dari sufiks *-i.*

3) Infiks: Merupakan afiks yang diletakkan di dalam kata dasar [12]. E.g.: -le (gelembung), -em(gemetar), -er (gerigi).

4) Simulfiks: Simulfiks adalah afiks yang menggantikan huruf di depan suatu kata. Afiks ini berfungsi untuk membentuk kata kerja dari suati kata dasar [12]. E.g.: *soto* (Indonesian food) => *nyoto (*eating soto*), sate* (Indonesian food) => *nyate* (eating sate).

5) Konfiks: Konfiks merupakan kombinasi dari prefiks dan sufiks untuk membentuk kata baru yang berasal dari kata dasar [1]. E.g.: *meng- -kan (menggunakan), di- per- -kan (dipertemukan).*

C. Levenshtein Distance

Salah satu metode yang digunakan untuk menentukan tingkat kemiripan antar string adalah *Levenshtein Distance*. Metode ini merupakan metode yang klasik dan mudah digunakan [research on]. Cara kerja dari metode ini adalah dengan menghitung seberapa banyak langkah yang dibutuhkan oleh string A untuk bisa menjadi string B. Semakin sedikit langkah yang dibutuhkan, maka kedua string tersebut semakin memiliki tingkat kemiripan yang tinggi [13]. Dapat dicontohkan kemiripan antara kata "Teman" dan "Temen", maka dari "Teman" untuk menjadi "Teman" dibutuhkan 1 perubahan huruf "e" pada huruf urutan ke. Sehingga jarak kemiripan antara "Teman" dengan "Teman" adalah 1. Contoh berhitungan *Levenshtein* dapat dilihat pada Tabel 2.

D. Data Kata Berimbuhan Tidak Baku

Dalam Bahasa Indonesia selain bahasa baku terdapat juga bahasa yang tidak baku yang sering digunakan dalam percakapan tidak resmi. Bahasa Indonesia memiliki variasai pemakaian dalam situasi tertentu, yaitu resmi dan tidak resmi [14]. Berdasarkan penelitian Zen [14] terdapat beberapa kata berimbuhan tidak baku. Data kata berimbuhan tidak baku diambil dari penelitian sebelumnya [2], berdasar pada penelitian Zen [14] yang dapat dilihat pada Tabel 1. Kata-kata berimbuhan tidak baku masing-masing memiliki fungsi dalam membentuk kata kerja, kata benda, dan kata sifat [14].

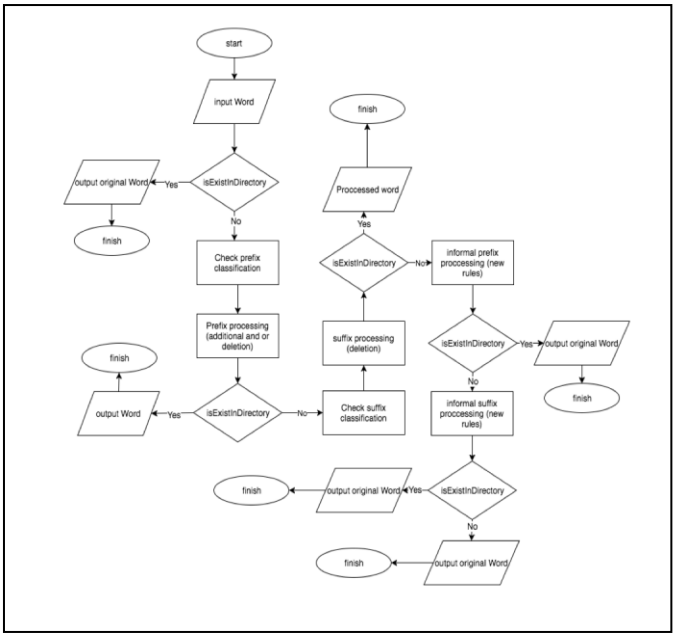
TABEL I
MATRIKS PERHITUNGAN JARAK LEVENHSTEIN

		t	**e**	**m**	**a**	**n**
	0	1	2	3	4	5
t	1	**0**	1	2	3	4
e	2	1	**0**	1	2	3
m	4	2	1	**0**	1	2
e	5	3	2	1	1	2
n	6	4	3	2	2	**0**

TABEL II
DAFTAR KATA BERIMBUHAN TIDAK BAKU

Bentuk dasar	**Bentuk tidak baku**	**Stemming Informal Affix**
Terjang	Nerjang	**Terjang**
Tuduh	Nuduh	**Tuduh**
Terima	Nerima	**Terima**
Tegur	negur	**Tegur**
Pukul	mukul	**Pukul**
Pimpin	Mimpin	**Pimpin**
Coba	nyoba	**Coba**
Siram	nyiram	**Siram**
Suruh	Nyuruh	**Suruh**
Simpan	Nyimpen	Nyimpen
Sebrang	Nyebrang	Nyebrang
Anggap	Nganggep	Nganggep
Amuk	Ngamuk	**Amuk**
Ambil	Ngambil	**Ambil**
Buka	Ngebuka	**Buka**
Bantu	Ngebantu	**Bantu**
Lepas	Ngelepas	**Lepas**
Bayang	Kebayang	**Bayang**
Injak	Keinjek	Keinjek
Kabul	Kekabul	**Kabul**
Pergok	Kepergok	**Pergok**
Tipu	Ketipu	**Tipu**
Ulang	Keulang	**Ulang**
Wujud	Kewujud	**Wujud**
Crita	Critain	Critain
Betul	Betulin	**Betul**
Manja	Manjain	**Manja**
Ganggu	Gangguin	**Ganggu**
Ganti	Gantian	**Ganti**
Ikut	Ikutan	**Ikut**
Musuh	Musuhan	**Musuh**
Sabun	Sabunan	**Sabun**
Teman	Temenan	Temenan

Tukar	Tukeran	Tukeran
Tanya	nanyain	**Tanya**
Tunjuk	nunjukin	**Tunjuk**
penting	mentingin	Ting
Pegang	megangin	**Pegang**
Selamat	nyelametin	Nyelametin
Sempat	nyempetin	nyempetin
Korban	ngorbanin	Ngorbanin
Hadap	ngadepin	Ngadepin
Bukti	ngebuktiin	**Bukti**
Warna	ngewarnain	**Warna**
Dengar	Kedengeran	Kedengeran
Ketemu	ketemuan	**Ketemu**
Benar	beneran	Beneran
Begini	ginian	Ginian
Kawin	kawinan	**Kawin**
Main	mainan	**Main**
Pargir	parkiran	**Parkir**
Dulu	duluan	**Dulu**
Gendut	gendutan	**Gendut**
Karat	karatan	**Karat**
Paling	palingan	**Paling**
Sabar	sabaran	**Sabar**
Bagus	kebagusan	**Bagus**
Sana	sanaan	**Sana**
Cepat	cepetan	Cepetan
pagi	sepagian	**Pagi**



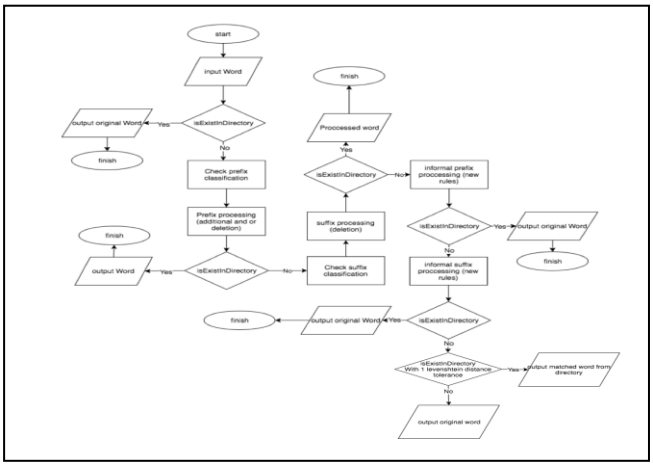
Gbr. 3 Algoritma dari *Non-formal Affix*

E. Modifikasi Algoritma dengan Levenshtein Distance

Pada penelitian sebelumnya yang berjudul "Non-formal Affixed Word Stemming in Indonesian Language" oleh Rahardyan Bisma dan Ema Utami [2], dilakukan modifikasi algorima Nazief & Andriani yang sudah ditingkatkan menggunakan *flexible affix classification* oleh Reina Setiawan [1] dengan penambahan aturan imbuhan tidak baku. Berikut dapat dilihat alur algoritma dari *non-forma affix* pada Gbr. 3.

TABEL III
KATA BERIMBUHAN TIDAK BAKU YANG TIDAK DAPAT DI-STEMMING PADA ALGORITMA NON-FORMAL AFFIX

Bentuk dasar	Bentuk tidak baku	Stemming Informal Affix
Simpan	Nyimpen	**Simpen => Nyimpen**
Seberang	Nyebrang	**Sebrang => Nyebrang**
Anggap	Nganggep	**Anggep => Nganggep**
Injak	Keinjek	**Injek => Keinjek**
Crita	Critain	**Crita => Critain**
Teman	Temenan	**Temen => Temenan**
Tukar	Tukeran	**Tuker => Tukeran**
penting	mentingin	Ting
Selamat	nyelametin	**Selamet => Nyelametin**
Sempat	nyempetin	**Sempet => nyempetin**
Korban	ngorbanin	Ngorbanin
Hadap	ngadepin	**Hadep => Ngadepin**
Dengar	Kedengeran	**Denger => Kedengeran**
Benar	beneran	**Bener => Beneran**
Begini	ginian	Ginian
Cepat	cepetan	**Cepet => Cepetan**



Gbr. 4 Algoritma Non-formal affix dengan *Levenshtein*

Dengan metode yang diusulkan, maka alur algoritma dimodifikasi dengan pendekatan *Levenshtein Distance*. Pengecekan kemiripan antar string dilakukan setelah seluruh proses pada algoritma *Non-formal Affix* selesai melakukan proses *stemming* dan gagal. Ketika *Non-formal Affix* gagal dalam melakukan *stemming* kata berimbuhan tidak baku, maka dilakukan pengecekan kemiripan antara kata yang

sudah dioleh oleh algoritma *Non-formal Affix* dengan kata yang berada dalam kamus menggunakan *Levenshtein Distance*. Pada algoritma *Non-formal Affix* terdapat beberapa kata yang tidak berhasil di *stemming* yang dapat dilihat pada Tabel 3.

Dari 16 kata yang tidak berhasil di *stemming* 13 di antaranya tidak dapat di-*stemming* karena jarak kemiripan antara kata dasar yang seharusnya dengan kata dasar hasil *stemming* adalah 1 (penghilangan, penambahan atau penggantian 1 karakter pada kata). Karena mengalami perubahan maka kata dasar hasil dari *stemming* tidak ditemukan dalam kamus kata dasar. Dari data pada Tabel 3 maka pada metode yang diusulkan kata hasil *stemming* akan dicocokan dengan kata pada kamus dengan toleransi jarak kemiripan yaitu 1. Sehingga algoritmanya terlihat seperti pada Gbr. 4.

F. Evaluasi

Kedua algoritma diimplemetasikan pada bahasa pemrogramman Jawa. Kata berimbuhan tidak baku dimasukkan pada *Array String*. Kemudian setiap kata pada array di-*stemming* satu per satu menggunakan perulangan. Hasil *stemming* dicetak pada *console log* yang kemudian di masukkan pada tabel perbandingan seperti pada Tabel 4.

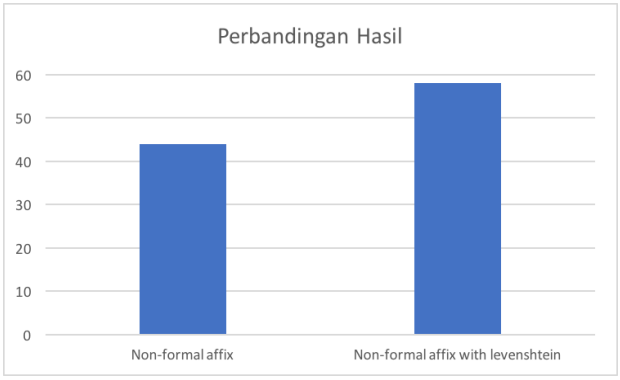
TABEL IV
HASIL STMMING ALGORITMA NON-FORMAL AFFIX DAN NON-FORMAL AFFIX DENGAN LEVENSHTEIN

Kata Dasar	Bentuk tidak baku	Non-formal Affix Levenshtein Algorithm	Non-formal Affix Algorithm
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**	Nyimpen
Sebrang	Nyebrang	**Seberang**	Nyebrang
Anggap	Nganggep	**Anggap**	Nganggep
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**	Keinjek
Kabul	Kekabul	**Kabul**	**Kabul**
Pergok	Kepergok	**Pergok**	**Pergok**
Tipu	Ketipu	**Tipu**	**Tipu**
Ulang	Keulang	**Ulang**	**Ulang**
Wujud	Kewujud	**Wujud**	**Wujud**
Crita	Critain	**Cerita**	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	**Teman**	Temenan
Tukar	Tukeran	**Tukar**	Tukeran
Tanya	nanyain	**Tanya**	**Tanya**
Tunjuk	nunjukin	**Tunjuk**	**Tunjuk**
penting	mentingin	Ting	Ting
Pegang	megangin	**Pegang**	**Pegang**
Selamat	nyelametin	**Selamat**	Nyelametin
Sempat	nyempetin	**Sempat**	nyempetin
Korban	ngorbanin	**Korban**	Ngorbanin
Hadap	ngadepin	**Hadap**	Ngadepin
Bukti	ngebuktiin	**Bukti**	**Bukti**
Warna	ngewarnain	**Warna**	**Warna**
Dengar	Kedengeran	**Dengar**	Kedengeran
Ketemu	ketemuan	**Ketemu**	**Ketemu**
Benar	beneran	**Benar**	Beneran
Begini	ginian	Ginian	Ginian
Kawin	kawinan	**Kawin**	**Kawin**
Main	mainan	**Main**	**Main**
Pargir	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	**Cepat**	Cepetan
pagi	sepagian	**Pagi**	**Pagi**

IV. HASIL DAN PEMBAHASAN

Pada penelitian ini menghasilkan algoritma *stemming* baru yang berdasar pada algoritma Nazief & Andriani yang sudah dimodifikasi dengan *Non-formal affix rule* yang ditujukan untuk melakukan *stemming* pada kata berimbuhan tidak baku. Dari hasil evaluasi maka diketahui tingkat akurasi dari algoritma *Non-formal Affix* dalam melakukan *stemming* kata berimbuhan tidak baku adalah **73.3 % atau 44 kata berhasil di *stemming* (0 *understemming*, 1 *overstemming*, 15 *unstemmed*)**.

Dengan ditambahkannya pendekatan *similarity* menggunakan metode *Levenshtein Distance* diketahui bahwa tingkat akurasi meningkat menjadi **96.6% atau 58 kata berhasil di-*stemming* (0 understemming, 1 overstemming, 0 unstemmed)**. Perbandingan banyak kata yang berhasil dilakukan *stemming* dari *Non-formal Affix* dan *Non-formal Affix with levenshtein* dapat dilihat pada Gbr. 5.



Gbr. 5 Perbandingan hasil *stemming*

V. KESIMPULAN

Dari hasil evaluasi yang sudah dilakukan dapat dilihat bahwa pendekata dengan menggunakan *Levenshtein Distance* dapat meningkatkan akurasi algoritma *stemming Non-formal Affix* dalam melakukan *stemming* kata berimbuhan tidak baku. Dengan pendekatan *Levenshtein Distance* masih belum dapat melakukan *stemming* pada 60 kata berimbuhan tidak baku. Kata yang tidak berhasil di-*stemming* adalah "ting" dikarenakan terdapat kata dasar "ting" pada *dictionary* dan kata "Ginian" dikarenakan belum terdapat *rule* yang untuk melakukan *stemming* kata tersebut.

Pada penelitian ini masih belum dilakukan evaluasi untuk *stemming* kata formal di luar 60 kata tidak baku dari data set penelitian ini. Topik ini masih memiliki peluang untuk dilakukan pengembangan kedepannya dari sisi kecepatan maupun efektifitas algoritma.

DAFTAR PUSTAKA

[1] R. Setiawan, A. Kurniawan, W. Budiharto, I. H. Kartowisastro and H. Prabowo, "Flexible affix classification for stemming Indonesian Language," *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Chiang Mai, 2016, pp. 1-6.

[2] Rahardyan Bisma, Ema Utami, "Non-formal Affixed Word Stemming in Indonesian Language," 2018 *International Conference on Information and Communication Technology (ICOIACT), Yogyakarta, 2018*.

[3] Mardiana Tari, Bharata Teguh, Hidayah Indriana, "Stemming Influence on Similiarity Detection of Abstract Written in Indonesia", in TELKOMNIKA, vol. 14, 2016, pp. 219-227.

[4] A. Aulia, D. Khairani and N. Hakiem, "Development of a retrieval system for Al Hadith in Bahasa (case study: Hadith Bukhari)," *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Denpasar, 2017, pp. 1-5.

[5] A. Sinaga, Adiwijaya and H. Nugroho, "Development of word-based text compression algorithm for Indonesian language document," *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, Nusa Dua, 2015, pp. 450-454.

[6] M. K. Keleş and S. A. Özel, "Similarity detection between Turkish text documents with distance metrics," *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, 2017, pp. 316-321.

[7] S. Zhang, Y. Hu and G. Bian, "Research on string similarity algorithm based on Levenshtein Distance," *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, 2017, pp. 2247-2251.

[8] A. Ene and A. Ene, "An application of Levenshtein algorithm in vocabulary learning," *2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, Targoviste, 2017, pp. 1-4.

[9] A. Aulia, D. Khairani and N. Hakiem, "Development of a retrieval system for Al Hadith in Bahasa (case study: Hadith Bukhari)," *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Denpasar, 2017, pp. 1-5.

[10] J. Asian, H. Williams dan S. Tahaghoghi, "Stemming Indonesian", in Conferences in Research and Practice in Information Technology Series, vol. 38, 2005, pp. 307-314.

[11] A.Z. Arifin, Mahendra and Ciptaningtyas, "Enchanced confix stripping stemmer and ants algorithm for classifying news document in Indonesian language".

[12] Khotimah Khusnul, "Analysis of Indonesian Affixes in English Words Found in Mobile Guide Edition: 54-59", in Thesis in English Departmen Faculty of Humanity Diponegoro University, 2012.

[13] D. Medhat, A. Hassan and C. Salama, "A hybrid cross-language name matching technique using novel modified Levenshtein Distance," *2015 Tenth International Conference on Computer Engineering & Systems (ICCES)*, Cairo, 2015, pp. 204-209.

[14] Zen Laily, "Non-formal Affix in Indonesian Informal Language Variety", in Lingua: Journal Ilmu Bahsaa dan Sastra, 2011.

[15] Emilya Ully Artha, Ahmad Dahlan, "Klasifikasi Model Percakapan Twitter Mengenai Ujian Nasional", in JPIT: Journal Pengembangan IT, 2018.