J-ICON, Vol. 9 No. 1, Maret 2021, pp. 55~63
DOI: 10.35508/jicon.v9i1.3749


# NAZIEF-ADRIANI STEMMER DENGAN IMBUHAN TAK BAKU PADA NORMALISASI BAHASA PERCAKAPAN DI MEDIA SOSIAL

**Katarina N. Lakonawa[1] , Sebastianus A. S. Mola[2] , Adriana Fanggidae[3]**

**[1,2,3]Program Studi Ilmu Komputer, Universitas Nusa Cendana, Jl. Adisucipto No. 10 Kupang Nusa Tenggara Timur**
**[1]Email: katarinalakonawa@gmail.com,**
**[2]Email: adimola@staf.undana.ac.id,**
**[3]Email: adrianafanggidae@staf.undana.ac.id**

## ABSTRAK

Penggunaan bahasa tak baku semakin marak dalam komunikasi di media sosial. Penggunaan bahasa tak baku tidak terbatas pada kalimat, klausa, atau frasa saja namun juga pada penggunaan kata. Pada penelitian ini, akan dilakukan normalisasi kata yang tak baku/ *nonstandard word* (NSW) tersebut ke kata baku/ *standard word* (SW) Bahasa Indonesia. Metode *stemmer* Nazief-Adriani (Nazief-Adriani *stemmer* (NAS)) dikembangkan menjadi *nonstandard stemmer* (NSS) dengan meningkatkan kemampuannya untuk mendeteksi imbuhan tak baku. Tujuan penelitian ini adalah membandingkan penggunaan NAS dan NSS dalam normalisasi NSW. Algoritma kemiripan Needleman-Wunsch digunakan untuk membobot hasil pencocokan. Hasil pengujian dengan *Mean Reciprocal Rank* (MRR) pada sebanyak 3.438 NSW didapatkan penggunaan NSS dengan jumlah kueri = 9 (Q=9) memiliki tertinggi sebesar 79.26% dengan rata-rata sebesar 50.48%. Sedangkan pengujian MRR menggunakan NAS dengan Q=9 mendapatkan hasil tertinggi sebesar 72.87% dan rata-rata sebesar 47.23%. Dari dua pengujian MRR yang dilakukan, ada 3 huruf yang memiliki hasil *stemming* tertinggi, baik dalam pengujian menggunakan NAS maupun menggunakan NSS yaitu huruf awal r, f dan j. Peningkatan nilai MRR paling signifikan terjadi pada huruf awal ‘d’, ‘n’ dan ‘t’ yang merupakan huruf awal dari sebagian imbuhan tak standar.
Kata kunci**:** kata tak baku, imbuhan tak baku, Nazief-Adriani *stemmer*, pencocokan *string* Needleman-Wunsch

## ABSTRACT

The use of non-standard language is increasingly prevalent in communication on social media. The use of indefinite language is not limited to sentences, clauses, or phrases but also word usage. In this study, the nonstandard word (NSW) will be normalized to the Indonesian standard word (SW). The Nazief-Adriani stemmer (NAS) method was developed into a nonstandard stemmer (NSS) by increasing its ability to detect non-standard additives. The Needleman-Wunsch similarity algorithm is used to weight the matches. The test results with the Mean Reciprocal Rank (MRR) of 3,438 NSW found that the use of NSS with the number of queries = 9 (Q = 9) had the highest of 79.26% with an average of 50.48%. Meanwhile, MRR testing using NAS with Q = 9 got the highest result of 72.87% and an average of 47.23%. Of the two MRR tests carried out, there were 3 letters that had the highest stemming results, both in tests using NAS and using NSS, namely the initial letters r, f and j. The most significant increase in MRR value occurs in the initial letters 'd', 'n' and 't' which are the initial letters of some non-standard affixes.
Keywords**:** nonstandard word, nonstandard affixes, Nazief-Adriani stemmer, Needleman-Wunsch string matching

## 1. PENDAHULUAN

Bahasa adalah sarana yang digunakan untuk berkomunikasi atau menyampaikan sesuatu dengan sesama. Dengan bahasa, manusia dapat berkomunikasi dengan sesama baik secara lisan maupun tulisan. Di berbagai kalangan baik yang muda maupun yang sudah tua, media sosial merupakan salah satu media yang saat ini sangat akrab. Kehadiran media sosial membawa dampak tersendiri terhadap penggunaan kata-kata bahasa percakapan. Bahasa Indonesia yang merupakan bahasa nasional bangsa Indonesia menjadi luntur karena banyaknya muncul bahasa percakapan yang tidak baku baik dalam struktur kalimat maupun kata-kata yang digunakan. Banyak orang menganggap bahasa ini lebih mudah dipakai dan dimengerti dalam berkomunikasi di media sosial dibandingkan bahasa Indonesia yang baku. Bahasa percakapan yang tidak baku ini perlu dinormalisasi ke bahasa Indonesia yang baku agar membantu masyarakat dalam


ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

berkomunikasi di lingkungan yang formal. Penelitian ini difokuskan pada normalisasi penggunaan kata yang tidak baku/ *nonstandard word* (NSW) menjadi kata baku *standard word* (SW).

Proses normalisasi NSW menjadi SW ini melalui tahapan beberapa tahapan seperti normalisasi angka ke huruf, normalisasi kata ulang, normalisasi *flooding*, dan *stemming* atau pemisahan kata dasar dari imbuhan. Setelah kata dasar ditemukan, maka selanjutnya akan dilanjutkan dengan proses pencocokan *string*. Pada penelitian ini, NSW yang dinormalisasi SW merupakan kata dasar (*root word*), yang dimana hanya, bukan frasa, klausa atau kalimat.

*Stemming* merupakan proses dalam sistem temu kembali informasi (*information retrieval*) yang mengubah kata-kata dalam sebuah dokumen ke kata dasar, sesuai dengan aturan-aturan tertentu yang digunakan. Sebagai contoh, kata 'bermain' akan ditransformasikan ke kata dasarnya yaitu 'main'. Untuk mendapatkan kata dasar dalam teks bahasa Indonesia maka prefix, sufiks dan konfiks perlu dihilangkan [1]. Ada beberapa algoritma yang digunakan untuk melakukan proses *stemming*, salah satunya algoritma Nazief-Adriani [2]. Ada beberapa peneliti menggunakan algoritma *stemming* Nazief-Adriani karena keakurtannya yang baik seperti dalam [2] yang menunjukan tingkat kesalahan 5%. Adapun penelitian [3] mengukur waktu komputasi untuk mencari kesamaan kata pada judul tulisan.

Setelah kata dasar diperoleh maka proses selanjutnya yang akan dilakukan yaitu pencocokan *string* dengan menggunakan algoritma Needleman-Wunsch. Pencocokan *string* adalah bagian terpenting dari proses pencarian *string* (*string searching*) pada suatu dokumen. Pengunaan teknik atau cara pencocokan *string* dapat menentukan hasil dari pencarian sebuah *string* dalam dokumen [4]. Ada beberapa algoritma pencocokan *string* yang biasa digunakan, salah satunya yaitu algoritma Needleman-Wunsch [5]. Penelitian [6] yang membandingkan algoritma Needleman-Wusch dan algoritma Lempel-Ziv dalam teknik global *sequence alignment* menunukan bahwa algoritma Needleman-Wunsh lebih unggul dalam hal kecepatan dibandingkan dengan algoritma Lempel-Ziv baik terutama untuk *dataset* yang besar.

Penelitian mengenai normalisasi NSW menjadi SW telah dilakukan dengan menerapkan berbagai metode. Metode konvensional berbasis leksikal telah dilakukan oleh [7] dengan terlebih dahulu mengidentifikasi jenis bahasa yang digunakan dan oleh [8] yang dikhususkan untuk *microtext*. Selanjutnya metode *nearest neighbour* digunakan dalam [9] untuk meningkatkan akurasi mormalisasi NSW. Dalam [10] telah dilakukan normalisasi menggunakan *finite state transducer* dengan penggunaan model 2-gram, 3-gram dan 4-gram dengan akurasi yang cukup baik (sekitar 78%).

Secara khusus, NSW dalam bahasa Indonesia selain dapat mengandung kata dasar yang tak baku juga dapat mengandung imbuhan tak baku seperti kata 'mnjanjikn' terdiri dari imbuhan 'me-kan' dengan kata dasar 'janji' atau kata 'membuwat' yang terdiri dari imbuhan awalan 'me' dan kata dasar 'buat'. Adanya imbuhan tak baku dalam kata belum dapat dideteksi oleh NAS [1]. Penelitian ini menawarkan modifikasi metode NAS dengan mengakomodir berbagai bentuk imbuhan tak baku yang sering digunakan di media sosial.

## 2. MATERI DAN METODE

### 2.1 Sumber dan Jenis Data

Sumber data bahasa percakapan di media sosial dalam penelitian diperoleh dari *https://github.com/nasalsabila/kamus-alay* [11]. Sedangkan sumber data untuk Kamus Besar Bahasa Indonesia (KBBI) didapat dari *https://github.com/kangfend/bahasa/tree/master/bahasa/data.* Jenis data yang digunakan dalam penelitian ini yaitu data kualitatif, di mana data kualitatif ini berupa NSW sebanyak 3.438 kata dan SW dari KBBI sebanyak 28.526 kata.

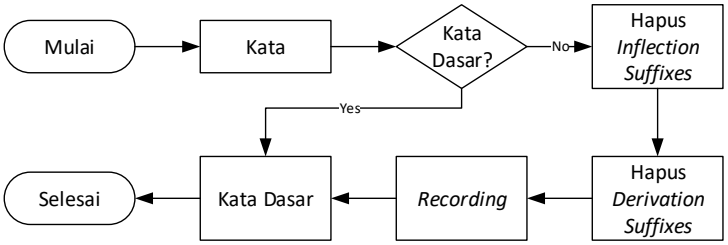
### 2.2 Algoritma Nazief-Adriani *Stemmer* (NAS)

Algoritma NAS merupakan algoritma yang berdasarkan pada aturan morfologi bahasa Indonesia, dan digabung menjadi satu kelompok lalu dienkapsulasi pada imbuhan yang dibolehkan (*allowed affixes*) dan imbuhan yang tidak dibolehkan (*disallowed affixes*). Kamus kata dasar dan *recoding* digunakan dalam algoritma ini, yakni menyusun kembali kata-kata yang mengalami proses *stemming* berlebih [2]. Secara umum, algoritma NAS ditampilkan pada gambar 1.

Algoritma NAS mempunyai langkah-langkah seperti berikut [1]:

1) Mencari kata yang akan distem dalam kamus. Apabila kata tersebut ditemukan maka dapat diasumsikan bahwa kata itu adalah kata dasar.Maka algoritma berhenti.
2) Menghapus *inflection suffixes* ("-lah", "-kah", "-ku", "-mu", atau "-nya"). Apabila berupa partikel/ ("-lah", "-kah", "-tah", atau "-pun") maka ulangi langkah ini untuk menghapus kata ganti milik ("-ku", "-mu", atau "-nya"), jika ada.


ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)




Gambar 1. Algoritma NAS

3) Hapus *derivation suffixes* (“-i”, “-an”, atau “-kan”). Apabila kata tersebut terdapat di kamus, maka algoritma berhenti. Tetapi jika tidak, maka lanjut ke langkah 3a.
   a) Jika “-an” sudah dihapus dan pada kata tersebut memiliki huruf terakhir “-k”, maka “-k” juga dihapus. Tetapi jika kata tersebut ada di kamus maka algoritma berhenti. Apabila tidak ada maka lanjut langkah 3b.
   b) Akhiran yang dihapus (“-i”, “-an”, atau “-kan”) dikembalikan, dan lanjut langkah 4.
4) Hapus *derivation prefix*. Apabila di langkah 3 ada imbuhan yang dihapus maka lanjut ke langkah 4a, tetapi jika tidak maka lanjut ke langkah 4b.
   a) Apabila ditemukan kombinasi awalan dan akhiran yang tidak diijinkan deperti pada tabel 1 maka algoritma berhenti, tetapi jika tidak maka lanjut ke langkah 4b.

Tabel 1. Kombinasi awalan dan akhiran yang tidak diijinkan

| Awalan | Akhiran yang tidak diijinkan |
|---|---|
| be- | -i |
| di- | -an |
| ke- | -i, -kan |
| me- | -an |
| se- | -i, -kan |
| te- | -an |

   b) Ulangi langkah 1 sampai 3 menentukan tipe awalan lalu hapus awalan. Apabila kata dasar belum ditemukan maka lakukan langkah 5, tetapi apabila sudah maka algoritma berhenti. Catatan: apabila awalan pertama sama dengan awalan kedua maka algoritma berhenti.
5) Melakukan *recoding*.
6) Apabila semua langkah sudah selesai namun tidak berhasil maka kata awal dapat diasumsikan sebagai kata dasar. Proses selesai.

### 2.3 Normalisasi Angka

Pada proses normalisasi angka, kata yang dimasukkan akan dicek apabila mengandung angka ‘1’, ‘2’,‘3’, ‘4’, ‘5’, ‘6’, ‘7’, ‘8’, ‘9’, atau ‘0’, maka angka tersebut akan diganti seperti pada tabel 2. Namun untuk angka ‘2’ tidak termasuk dalam proses normalisai angka ke huruf karena angka ‘2’ dibatasi untuk proses normalisasi kata ulang. Contohnya kata “ap4” maka angka ‘4’ akan diganti menjadi huruf ‘a’ sehingga kata tersebut menjadi “apa”. Normalisasi angka ke sebutan angka terjadi jika angka tersebut merupakan satu kata tersendiri atau posisinya berada setelah awalan. Contohnya kata ‘ber3’ akan diubah menjadi ‘bertiga’.

### 2.4 Normalisasi Kata Ulang

Normalisasi kata ulang yang dimaksud yaitu jika pada kata yang dimasukkan terdapat angka ‘2’, tanda /’/ dan /”/, maka angka atau tanda tersebut akan dinormalisasi dengan menambah tanda “-”. Contohnya kata “mana2” maka akan dinormalisasi menjadi “mana-mana”.

### 2.5 Normalisasi *Flooding*

Normalisasi *flooding* yang dimaksud adalah menhapus huruf yang berulang yang terdapat pada NSW. Apabila terdapat 2 atau lebih huruf ‘c’, ‘f’, ‘h’, ‘j’, ‘p’, ‘q’, ‘r’, ‘u’, ‘v’, ‘w’, ‘x’, ‘y’, dan ‘z’ maka


ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

akan dihapus perulangannya sehingga hanya satu huruf saja yang dipertahankan. Demikian juga jika terdapat deretan huruf yang berulang seperti 'hahaha' maka akan diubah menjadi 'haha'.

Tabel 2. Normalisasi angka ke huruf

| Angka | Normalisasi ke huruf | Normalisasi ke sebutan angka |
|---|---|---|
| 0 | o | nol |
| 1 | l | satu |
| 2 | - | dua |
| 3 | e | tiga |
| 4 | a | empat |
| 5 | s | lima |
| 6 | g | enam |
| 7 | j | tujuh |
| 8 | b | delapan |
| 9 | 9 | sembilan |

### 2.6 Modifikasi Algoritma *Non-Standard Stemmer* (NSS) Berdasarkan NAS

Algoritma NSS pada dasarnya sama dengan NAS namun NSS meningkatkan kemampuan NAS dalam *stemming* NSW dengan menambahkan beberapa varian imbuhan, partikel dan kata ganti milik tak baku. Beberapa bentuk imbuhan tak baku yang dapat dideteksi adalah:

1) Kata ganti milik tak baku ("-nyah", "-ny", "-nye").
2) Partikel tak baku ("-kh", "-lh", "-th", "-pn").
3) Imbuhan awalan tak baku ("n-", "m-", "ng-", "ny-").
4) Imbuhan akhiran tak baku ("-kn", "-in").
5) Imbuhan awalan tak baku ("d-", "k-", "s-", "t-", "t-", "m-", "p-").
6) Imbuhan awalan tak baku ("br-", "bl-").
7) Imbuhan awalan tak baku ("te-", "tr-").
8) Imbuhan awalan tak baku ("mm-", "mn-", "mng-", "mg-", "mny-", "my-").
9) Imbuhan awalan tak baku ("pr-", "pl-", "pm-", "png-", "pg-", "pny-", "py-").
10) Semua *derivation prefix* yang merupakan kombinasi dari imbuhan awalan dan akhiran baku dan tak baku.

### 2.4 Algoritma Pencocokan *String* Needleman-Wunsch

Pengunaan teknik atau cara pencocokan string dapat menentukan hasil dari pencarian sebuah *string* dalam dokumen [4]. Algoritma Needleman-Wuncsh merupakan salah satu implementasi program dinamis dari perluasan dari algoritma pencocokan *string* pada barisan atau teks. Algoritma ini biasanya digunakan untuk menentukan tingkat kecocokan atau kesamaan dua buah teks [12]. Langkah-langkah algoritma Needleman-Wunsch yang digunakan pada penelitian ini adalah [12] :

1) Misalkan kata *A* adalah NSW dari proses *stemming* akan dicocokan dengan kata *B* yang adalah SW pada KBBI. Pada langkah ini, akan dilakukan dengan mencari semua kata yang ada di dalam KBBI yang mempunyai huruf pertama sama dengan huruf pertama kata *A*.
2) Inisialisasi matriks
   a) Menentukan nilai kecocokan, ketidakcocokan dan pinalti (*gap*). Untuk pemberian nilai bisa bermacam-macam tergantung dari defenisi penelitian yang diinginkan.
   b) Membuat matriks berukuran :
      Jumlah baris = panjang kata *A* + 1
      Jumlah kolom = panjang kata *B* + 1
   c) Menentukan elemen matriks pada baris ke-*i* dan kolom ke-*j* :
      $F_{i0}$ = 0 untuk *i* = 1 sampai dengan panjang kata *A*
      $F_{0j}$ = 0 untuk *j* = 1 sampai dengan panjang kata *B*
3) Mengisi semua elemen matriks pada baris ke-1 sampai dengan baris panjang kata *A*. Pengisian nilai matriks tiap sel dilakukan dengan aturan seperti pada persamaan 1.

$$F_i = \max \begin{cases} F_{i-1,j-1} + S(A_i B_j) & (diagonal\ score) \\ F_{i,j-1} + d & (left\ score) \\ F_{i-1,j} + d & (up\ score) \end{cases} \quad \ldots\ldots (1)$$

di mana,

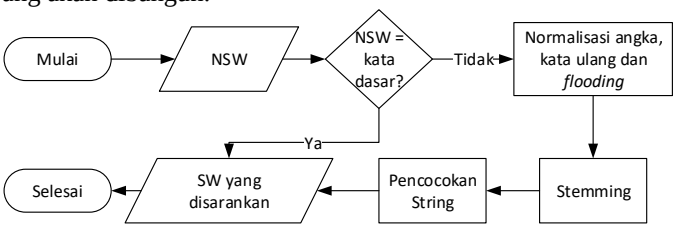

ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

$S(A_i B_j)$ : nilai ketidakcocokan / kecocokan antara karakter ke-*i* teks *A* dan karakter ke-*j* teks *B*
*d* : nilai pinalti atau gap

4) *Trace back*
   a) Melakukan *traceback* mulai dari sel kanan bawah sampai ke sel kiri atas dengan mengikuti arah tanda panah. Urutan jalur dibangun berdasarkan aturan sebagai berikut :
      - Panah diagonal menunjukkan kecocokan atau ketidakcocokan, sehingga huruf kolom dan huruf baris sel sejajar.
      - Panah horizontal atau vertikal mewakili gap. Panah horizontal akan mensejajarkan celah ('-') dengan huruf pada baris, dan panah vertikal akan mensejajarkan celah ('-') dengan huruf pada kolom.
   b) Langkah terakhir yaitu mencocokan kedua teks tersebut sesuai dengan arah jalur lalu menghitung skor.

### 2.5 Rancangan Sistem

Rancangan sistem bertujuan untuk mengetahui alur proses dalam sistem dan memberikan gambaran perancangan sistem yang akan dibangun.




Gambar 2. *Flowchart* sistem

Berdasarkan gambar 2 secara umum dapat dijelaskan sebagai berikut:

1) Langkah pertama yaitu memasukkan NSW dari *dataset* yang digunakan.
2) Setelah kata dimasukkan, sistem akan langsung mengecek apakah kata yang dimasukkan merupakan kata dasar sesuai dengan KBBI atau tidak. Apabila kata tersebut merupakan kata dasar sesuai KBBI maka sistem akan langsung menampilkan kata dan proses selesai. Tetapi, apabila kata tersebut bukan merupakan kata dasar sesuai KBBI maka akan dilanjutkan ketahap selanjutnya.
3) Pemrosesan awal NSW dengan melakukan normalisasi angka, kata ulang dan *flooding*.
4) Proses selanjutnya yaitu melakukan *stemming* Apabila kata yang dimasukkan mempunyai imbuhan, maka pada tahap ini imbuhan tersebut akan dihapus dan diambil kata dasarnya. Contohnya seperti kata "memakan", maka imbuhan "me-" akan dihapus dan diambil kata dasarnya yaitu "makan". Dalam proses ini untuk mengecek kata berimbuhan, terdapat dua macam *stemming* yang digunakan yaitu NAS dan NSS.
5) Langkah selanjutnya yaitu pencocokan *string* Needlman-Wunsch. Algoritma Needlman-Wunsch merupakan algoritma yang digunakan untuk menentukan tingkat kecocokan atau kesamaan dua buah teks.
6) Setelah hasil dari pencockan *string* didapat, sistem akan menampilkan kata dengan skor tertinggi dan proses selesai.

### 2.6 *First Character Match*

Pada penelitian ini, strategi pencocokan *string* menggunakan metode *first character match* (FCM). Metode ini melakukan pencocokan dengan membandingkan NSW dengan SW dari KBBI yang memiliki karakter awal yang sama dengan kata tidak baku.

### 2.7 *Mean Reciprocal Rank*

Untuk menguji keakuratan sistem dapat menggunakan metode *Mean Reciprocal Rank* (MRR). MRR merupakan suatu metode perhitungan statistik yang digunakan untuk menghitung tingkat kebenaran dari sebuah daftar jawaban. Nilai *reciprocal rank* (RR) disesuaikan dengan posisi jawaban pada daftar jawaban yang diberikan seperti pada persamaan 2 [13].

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad \text{..... (2)}$$


ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

di mana Q merupakan kueri dan $rank_i$ mengacu pada posisi peringkat pertama dokumen yang relevan untuk kueri ke-i.

### 2.8 Contoh Normalisasi NSW denganNSS

Misalkan terdapat sebuah NSW dari percakapan di media social 'mknlh'. Tahapan berikut ini akan menjelaskan proses normalisasi NSW tersebut menjadi SW berdasarkan *flowchart* pada gambar 1.

1) NSW: 'mknlh'
2) Pengecekan kata dasar. Berdasarkan penelusuran dari KBBI, kata 'mknlh' tidak ditemukan sehingga dapat dikatakan bahwa NSW tersebut bukan kata dasar.
3) Pemrosesan awal:
   a. Normalisasi angka: tidak ditemukannya angka dalam NSW
   b. Normalisasi kata ulang: NSW bukan kataulang
   c. Normalisasi *flooding*: tidak ada *flooding* dalam NSW
4) Proses *stemming*:
   a. Jika menggunakan NAS maka kata 'mknlh' akan dianggap sebagai kata dasar (tidak ditemukannya imbuhan).
   b. Jika menggunakan NSS maka kata 'mknlh' akan menjadi 'mkn' karena ditemukan akhiran partikel 'lah' dalambentuk tidak baku 'lh'.
5) Pencocokan *string* dengan algortima Needleman-Wunsch. Dalam tahapan ini akan dilakukan pencocokan NSW 'mkn'dengan semua kata dalam KBBI dengan huruf awal 'm'. Ilustrasi berikut ini menunjukan pencocokan NSW 'mkn' dengan kata 'makan' menggunakan algoritma Needleman Wunsch [12]:
   a. Inisialiasi matriks
      i. Menentukan nilai kecocokan, ketidakcocokan dan pinalti (*gap*). Misalkan digunakan nilai kecocokan = 4, ketidakcocokan = -3 dan gap = -1.
      ii. Membentuk matriks berukuran :
          Jumlah baris = panjang NSW + 1 = 4
          Jumlah kolom = panjang SW + 1 = 6
      iii. Menentukan elemen matriks pada baris ke-*i* dan kolom ke-*j* :
          $F_{i0}$ = 0 untuk *i* = 1 sampai dengan panjang NSW
          $F_{0j}$ = 0 untuk *j* = 1 sampai dengan panjang SW
   b. Pengisian matriks
      i. Hasil pengisian setiap elemen pada baris ke-0 dan kolom ke-0 dengan 0 seperti pada gambar 3.

| | | *j*=0 | *j*=1 | *j*=2 | *j*=3 | *j*=4 | *j*=5 |
|---|---|---|---|---|---|---|---|
| | | 0 | m | a | k | a | n |
| *i*=0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *i*=1 | m | 0 | | | | | |
| *i*=2 | k | 0 | | | | | |
| *i*=3 | n | 0 | | | | | |

Gambar 3. Hasil pengisian matriks baris ke-0 dan kolom ke-0

      ii. Pengisian baris ke-1 dan kolom ke-1. Karena karakter ke-1 NSW ('m') sama dengan karakter ke 1 SW ('m') maka $S_{11}$ = 4.
          - *Diagonal score* = $F_{00} + S_{11} = 0 + 4 = 4$
          - *Left score* = $F_{10} + d = 0 + (-1) = -1$
          - *Up score* = = $F_{01} + d = 0 + (-1) = -1$

          Oleh karena itu, $F_{11}$ = max ($F_{00} + S_{11}$, $F_{10} + d$, $F_{01} + d$) = max (4, -1, -1) = 4.
          Tanda panah pada gambar 4 menunjukan nilai 4 diperoleh dari *diagonal score.*
      iii. Pengisian matriks akan diteruskan hingga semua elemen yang kosong terisi. Matriks lengkapnya seperti pada gambar 5.
   c. *Trace back*
      Hasil *trace back* mengikuti panah dari pojok kanan bawah matriks menuju pojok kiri atas. Hasil *trace back* seperti pada gambar 6.


ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

| | | *j*=0 | *j*=1 | *j*=2 | *j*=3 | *j*=4 | *j*=5 |
|---|---|---|---|---|---|---|---|
| | | 0 | m | a | k | a | n |
| *i*=0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *i*=1 | m | 0 | 4 | | | | |
| *i*=2 | k | 0 | | | | | |
| *i*=3 | n | 0 | | | | | |

Gambar 4. Hasil pengisian matriks kolom ke-1 baris ke-1

| | | *j*=0 | *j*=1 | *j*=2 | *j*=3 | *j*=4 | *j*=5 |
|---|---|---|---|---|---|---|---|
| | | 0 | m | a | k | a | n |
| *i*=0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *i*=1 | m | 0 | 4 | 3 | 2 | 1 | 0 |
| *i*=2 | k | 0 | -1 | 2 | 7 | 6 | 5 |
| *i*=3 | n | 0 | -1 | 1 | 6 | 5 | 10 |

Gambar 5. Hasil pengirian matriks lengkap

| | | *j*=0 | *j*=1 | *j*=2 | *j*=3 | *j*=4 | *j*=5 |
|---|---|---|---|---|---|---|---|
| | | 0 | m | a | k | a | n |
| *i*=0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *i*=1 | m | 0 | 4 | 3 | 2 | 1 | 0 |
| *i*=2 | k | 0 | -1 | 2 | 7 | 6 | 5 |
| *i*=3 | n | 0 | -1 | 1 | 6 | 5 | 10 |

Gambar 6. *Trace back* pada matriks

d. Pencocokan kedua string dan menghitung skor. Hasil pencocokannya adalah:
NSW : m-k-n
SW : makan
Skor = 4 + (-1) + 4 + (-1) + 4 = 10

6) Perangkingan
Setelah dilakukan perhitungan kemiripan untuk membandingkan NSW dengan strategi penelusuran FCM maka yang dipilih sebagai kata yang disarankan adalah kata dengan skor terbesar.

7) Perhitungan akurasi dengan MRR
Untuk semua NSW dalam dataset akan dihitung *reciprocal rank* dan kemudian dihitung MRR. Misalkan selain NSW 'mknlh', terdapat 2 NSW lain yakni 'ank', dan 'dn'. Tabel 3 menunjukan peringkat berkebalikan dari ketiga NSW tersebut. Kata pada kolom hasil pencocokan yang ditulis tebal (*bold*) menunjukan posisi SW dalam peringkat 3 terbaik. Berdasarkan persamaan 2 dapat dihitung nilai MRR sebesar:

MRR = 1/3 * (1+1+1/3) = 0,77.

Tabel 3. Hasil perangkingan dengan MRR

| No | NSW | Hasil pencocokan untuk 3 terbaik (Q=3) | SW | Peringkat kata baku (*rank*) | Peringkat berkebalikan (*Reciprocal Rank*) |
|---|---|---|---|---|---|
| 1 | mknlh | **makan**, makin, makna | makan | 1 | 1 |
| 2 | ank | **anak**, antuk, antik | anak | 1 | 1 |
| 3 | dn | din, den, **dan** | dan | 3 | 1/3 |

## 3. HASIL DAN PEMBAHASAN

Pada penelitian ini, digunakan metode FCM sebagai strategi pencocokan *string* dan menetapkan nilai *gap* = -1, *miss* = -3 dan *match* = 4 untuk semua karakter. Pengujian dilakukan dengan mencocokan NSW yang ada di *dataset* dengan SW yang terdapat di KBBI menggunakan algoritma NAS dan NSS dengan algoritma pencocokan *string* Needleman-Wunsch. Ada dua pengujian yang dilakukan yaitu menggunakan NAS dan NSS. Hasil pengujian menggunakan algoritma NAS dan algoritma pencocokan


ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

*string* Needleman-Wunsch, bisa dilihat pada tabel 2 dan sedangkan tabel 3 menunjukan hasil pengujian dengan NSS dan Needleman-Wunsch.

Dari tabel 4 dan 5 diketahui bahwa penggunaan NSS dalam normalisasi NSW dapat meningkatkan rata-rata akurasi sistem. Dengan menggunakan NAS, nilai rata-rata MRR sebesar 47, 23% untuk Q=9 sedangkan dengan NSS meningkat menjadi 50,48%. Demikian pula dengan nilai maksimum MRR untuk NAS dengan Q=9 hanya sebesar 72,87% dan meningkat menjadi 79,26% jika menggunakan NSS. Hal ini menunjukan bahwa penggunaan NSS yang mengakomodir imbuhan tak baku sangat perlu dilakukan.

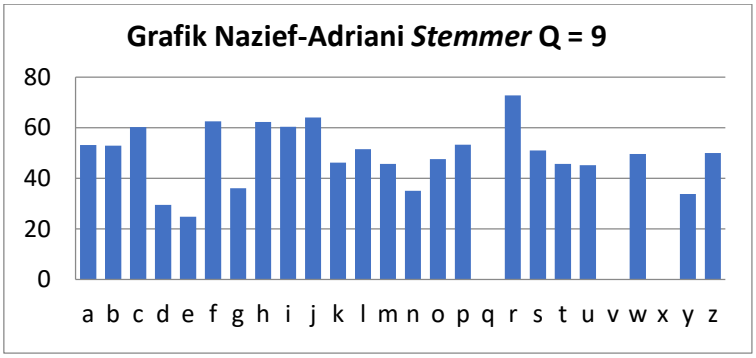
Tabel 4. Hasil Pengujian menggunakan NAS

| Pengujian | Nilai MRR (%) | | | |
|---|---|---|---|---|
| | Q = 3 | Q = 5 | Q = 7 | Q = 9 |
| Rata- rata | 45.8 | 46.77 | 47.03 | 47.23 |
| Nilai max | 71.76 | 72.87 | 72.87 | 72.87 |

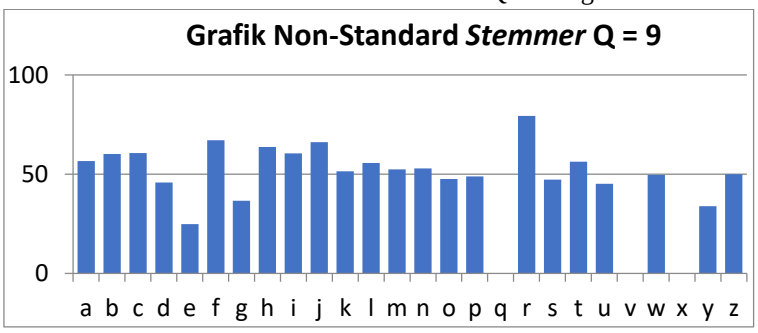
Tabel 5. Hasil Pengujian menggunakan NSS

| Pengujian | Nilai MRR (%) | | | |
|---|---|---|---|---|
| | Q = 3 | Q = 5 | Q = 7 | Q = 9 |
| Rata- rata | 49.14 | 50.05 | 50.3 | 50.48 |
| Nilai max | 78.7 | 79.26 | 79.26 | 79.26 |

Gambar 7 dan gambar 8 menunjukan grafik nilai rata-rata MRR per huruf awal NSW untuk Q = 9 dengan menggunakan NAS dan NSS. Kedua gambar tersebut secara konsisten menunjukan bahwa nilai rata-rata MRR tertinggi terdapat pada NSW dengan huruf awal ‘r’, ‘f’ dan ‘j’. Untuk huruf awal NSW ‘q’, ‘v’ dan ‘x’ nilai rata-rata MRR adalah 0 karena tidak ada NSW dengan huruf awal tersebut. Perbaikan akurasi juga sangat signifikan terlihat pada huruf awal NSW ‘d’, ‘n’, dan ‘t’ dengan menggunakan NSS. Hal ini disebabkan karena NSS mengakomodir imbuhan tak baku yang sebagian di antaranya memiliki huruf awal ‘d’, ‘n’ dan ‘t’.




Gambar 7. Nilai rata-rata MRR untuk Q=9 dengan NAS




Gambar 8. Nilai rata-rata MRR untuk Q=9 dengan NSS

Hasil dari penelitian ini menunjukan bahwa apabila dibandingkan, penggunaan NAS memberikan hasil yang lebih baik dibandingkan dengan NSS dalam normalisasi NSW.

## 4. KESIMPULAN DAN SARAN

Penggunaan NSS pada normalisasi NSW menjadi SW membawa peningkatan akurasi sistem yang ditandai dengan neningkatnya nilai rata-rata MRR dibandingkan dengan NAS. Peningkatan ini terjadi untuk


ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

semua kueri Q yang diujicobakan. Peningkatan rata-rata MRR terbaik pada Q = 9 sebesar 3,25% terjadi karena penggunaan NAS pada NSW dengan huruf awal 'r', 'f' dan 'j'. Walaupun peningkatan yang terjadi tidak cukup signifikan terhadap keseluruhan data namun peningkatan ini menunjukan adanyan 112 data dalam *dataset* NSW yang memiliki imbuhan non standar yang mampu dinormalisasi. Perbaikan paling signifikan terjadi pada huruf awal 'd', 'n' dan 't' yang merupakan huruf awal dari sebagian imbuhan tak standar. Penggunaan NSS secara keseluruhan menunjukan hasil yang lebih baik dari NSS.

Dengan adanya penelitian ini membuka peluang penelitian lanjutan dalam normalisasi NSW. Persoalan adanya kata serapan dari bahasa asing dan daerah, singkatan/ akronim, dan frasa tanpa spasi belum dapat diakomodir. Selanjutnya perlu juga dilakukan penelitian mengenai pembobotan kecocokan, ketidakcocokan dan *gap* secara dinamis. Lebih lanjut harus dilanksanakan penelitian dengan mengakomodir jenis kata dan posisinya di dalam kalimat.

## DAFTAR PUSTAKA

[1] L. Agusta, 'Perbandingan algoritma stemming Porter dengan algoritma Nazief & Adriani untuk stemming dokumen teks bahasa indonesia', *Konferensi Nasional Sistem dan Informatika*, vol. 2009, pp. 196–201, 2009.
[2] D. Wahyudi, T. Susyanto, and D. Nugroho, 'Implementasi dan analisis algoritma stemming nazief & adriani dan porter pada dokumen berbahasa indonesia', *Jurnal Ilmiah SINUS*, vol. 15, no. 2, Art. no. 2, 2017.
[3] M. W. Sardjono, M. Cahyanti, M. Mujahidin, and R. Arianty, 'Pendeteksi Kesamaan Kata untuk Judul Penulisan Berbahasa Indonesia Menggunakan Algoritma Stemming Nazief-Adriani', *Sebatik*, vol. 22, no. 2, Art. no. 2, 2018.
[4] M. A. Saragih, 'Implementasi Algoritma Brute Force dalam Pecncocokan Teks Font Italic Untuk Kata Berbahasa Inggris pada Dokumen Microsoft Office Word', *Pelita Informatika Budi Darma*, vol. 4, pp. 84–86, 2013.
[5] M. R. F. Zen, S. W. Putri, and M. F. Rasyid, *Penerapan Algoritma Needleman-Wunsch sebagai Salah Satu Implementasi Program Dinamis pada Pensejajaran DNA dan Protein*. Laboratorium Ilmu dan Rekayasa Komputasi, Program Studi Teknik Informatika, 2006.
[6] M. A. Malendes and H. Bunyamin, 'Analisa Perbandingan dan Implementasi Algoritma DNA Pairwise Sequence Alignment Needleman-Wunsch dan Lempel-Ziv', *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 3, no. 1, Art. no. 1, 2017.
[7] A. M. Barik, R. Mahendra, and M. Adriani, 'Normalization of Indonesian-English Code-Mixed Twitter Data', in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 2019, pp. 417–424.
[8] D. Gunawan, Z. Saniyah, and A. Hizriadi, 'Normalization of Abbreviation and Acronym on Microtext in Bahasa Indonesia by Using Dictionary-Based and Longest Common Subsequence (LCS)', *Procedia Computer Science*, vol. 161, pp. 553–559, 2019.
[9] S. A. Ansari, U. Zafar, and A. Karim, 'Improving text normalization by optimizing nearest neighbor matching', *arXiv preprint arXiv:1712.09518*, 2017.
[10] J. Porta and J.-L. Sancho, 'Word Normalization in Twitter Using Finite-state Transducers.', *Tweet-Norm@ SEPLN*, vol. 1086, pp. 49–53, 2013.
[11] N. A. Salsabila, 'nasalsabila/kamus-alay', Aug. 19, 2020. https://github.com/nasalsabila/kamus-alay (accessed Oct. 06, 2020).
[12] A. R. Dewi, 'Penerapan Algoritma Needleman-Wunsch untuk Mengidentifikasi Mutasi pada Sekuen DNA Virus Korona-Application Of Needleman-Wunsch Algorithm To Identify Mutations In Corona Virus DNA Sequences', PhD Thesis, Institut Teknologi Sepuluh Nopember, 2018.
[13] R. Sunartio, H. N. Palit, and A. Gunawan, 'Hotel Recommender System Menggunakan Metode Pendekatan Graph pada Dataset Trivago', *Jurnal Infra*, vol. 8, no. 1, Art. no. 1, 2020.


ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)