Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based

Anita Guterres
Magister Teknologi Infomasi
Sekolah Tinggi Teknik Surabaya
anitaguterres85@gmail.com

Gunawan
Magister Teknologi Informasi
Sekolah Tinggi Teknik Surabaya
gunawan@stts.edu

Joan Santoso
Magister Teknologi Infomasi
Sekolah Tinggi Teknik Surabaya
joan@stts.edu

Abstrak - *Stemming* adalah proses yang sangat penting untuk mencari kata dasar dari sebuah kata derivatif. Inti dari proses *stemming* adalah menghilangkan imbuhan pada suatu kata. *Stemming* sangat dibutuhkan untuk proses *information retrieval system*. Algoritma pada proses *stemming* bisa berbeda-beda pada setiap bahasa di berbeda negara. Data yang digunakan adalah 176 kata dasar dalam bahasa Tetun yang merupakan bahasa asli warga negara Timor Leste. Penelitian ini bertujuan untuk merancang algoritma baru yang tepat untuk *stemming* bahasa Tetun. Tahap awal *stemming* bahasa Tetun adalah proses filterisasi untuk menghilangkan tanda baca, angka, dan kata yang tidak penting. Lalu tahap tokenisasi untuk membuat variabel yang terdiri dari satu kata. Lalu setiap kata melalui proses *stemming* untuk menghilangkan imbuhan awalan, akhiran, dan konfiks. Analisis dilakukan berdasarkan kasus *error stemming* seperti *overstemming*, *understemming*, *unchanged*, dan *spelling exception*. Hasil uji coba yang didapatkan adalah algoritma *stemming* bahasa Tetun menghasilkan akurasi sebesar 90.52%.

Kata Kunci: Bahasa Tetun, *Stemmer*

I. LATAR BELAKANG

Stemming merupakan suatu proses untuk mengubah kata berimbuhan menjadi kata dasar. Pada umumnya bahasa Tetun merupakan bahasa resmi yang digunakan negara Timor Leste menjadi bahasa kerja. Namun bahasa Tetun sendiri ada 2 macam yaitu Tetun Frasa (Tetun Dili) dan Tetun Terik (Suai, Viqueque, dan Belu Nusa Tenggara Timur di Indonesia) yang terletak di perbatasan Indonesia dan Timor Leste. Di Belu, bahasa Tetun digunakan sebagai bahasa sehari hari dan lebih menyerap bahasa Indonesia karena terpisah dari negara Timor Leste setelah merdeka. Sedangkan Tetun Terik di Viqueque dan Suai digunakan sebagai bahasa kerja. Pada Tetun Frasa dan Tetun Terik masih terdapat kata yang belum ditemukan padanannya oleh ahli bahasa Tetun sehingga masih meminjam dari bahasa lain seperti bahasa Portugis dan bahasa Indonesia. Walaupun bahasa Tetun adalah bahasa resmi dan digunakan sebagai bahasa kerja, tetapi masih sulit diterapkan di tempat kerja, sehingga dalam prakteknya masih menggunakan bahasa campuran seperti bahasa Portugis dan bahasa Indonesia, dikarenakan 2 negara tersebut pernah menjajah Timor Leste.

II. TINJAUAN ALGORITMA STEMMER PADA BAHASA LAIN

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata dengan menghilangkan semua imbuhan (*affixes)* baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*), dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan. *Stemming* adalah *tools* dasar pemprosesan teks yang digunakan untuk *text retrieval* (Frakes, 1992), mesin penterjemah (Bakar dan Rahman, 2003), meringkas dokuman (Orasan *et al.*, 2004) dan klasifikasi teks (Gaustad ang Bouma, 2002). Berdasarkan asumsi bahwa *term* yang memiliki akar kata yang sama akan selalu memiliki makna yang sama, *stemming* digunakan pada *information retrieval* untuk meningkatkan keakurasian *retrieval* (perolehan informasi). Selain untuk meningkatkan keakurasian *retrieval*, *stemming* yang dilakukan pada proses *indexing* juga akan mengurangi ukuran dari indeks *file* [1].

A. Bahasa Indonesia

Algoritma Nazief dan Adriani *Stemmer* diciptakan oleh Bobby A. A. Nazief dan Mirna Adriani yang berasal dari Universitas Indonesia pada tahun 1996. Pada sub bab ini akan dijelaskan tentang algoritma dan contoh penggunaanya.

Algoritma ini sangat berhubungan dengan peraturan morfologi bahasa Indonesia dengan mengelompokkan dan meringkas imbuhan-imbuhan yang diperbolehkan. Imbuhan yang dimaksud adalah awalan, akhiran, gabungan antara awalan dengan akhiran, dan sisipan. Algoritma ini memiliki proses *recording* yang digunakan untuk mengembalikan huruf awal dari kata dasar yang hilang karena proses penghilangan akhiran dari sebuah kata. Algoritma ini juga memiliki kamus data kata dasar yang digunakan pada setiap langkah untuk memeriksa hasil proses *stemming* untuk mendapat kata dasar yang dicari [2].

1. Akhiran Infleksi (*Inflection Suffix*)
   Akhiran infleksi merupakan akhiran yang tidak mengubah kata dasarnya. Contohnya adalah kata "baca" diberi akhiran "-lah" maka akan membentuk kata "bacalah". Akhiran infleksi terdiri dari dua macam, yaitu:
   a. Kata sandang/*particles* (P)

Terdiri dari akhiran "-lah" dan "-kah". Contohnya adalah kata "bacalah", "benarkah".

b. Kata ganti milik/*Possessive Pronouns* (PP)
Terdiri dari akhiran "-ku", "-mu", "-nya". Contohnya adalah kata "ayahku", ayahmu", "ayahnya".

Kata sandang dan kata ganti milik bisa digunakan bersama dalam satu kata. Jika digunakan dalam satu kata maka kata ganti milik akan muncul sebelum kata sandang.

2. Akhiran Derivasi/Derivation Suffix (DS)
Akhiran Derivasi adalah kumpulan akhiran yang bisa digabungkan bersama-sama pada kata dasar. Setiap kata hanya bisa diberi lebih dari satu akhiran derivasi. Contohnya adalah kata "mati" bisa diberi akhiran "-kan" maka akan membentuk kata "matikan". Lalu diberi satu akhiran lagi dengan akhiran "-lah" maka akan membentuk kata "matikanlah".
3. Awalan Derivasi/*Derivation Prefix* (DP)
Awalan Derivasi adalah kumpulan awalan yang bisa digabungkan bersama-sama pada kata dasar, atau pada kata-kata yang punya maksimal dua awalan. Contohnya adalah kata "indah" bisa diberi awalan "mem-" dan "per-" sehingga akan membentuk kata "memperindah".

B. Bahasa India

Pada tahun 2010, Dinesh Kumar dan Pangeran Rana mengembangkan desain dan pengembangan *stemmer* untuk Punjabi, menggunakan algoritma *Brute Force* untuk membendung kata Punjabi. Pada tahun 2001, Shambhavi dkk. mengenalkan penganalisis morfologi Kannada. *Stemmer* untuk bahasa Hindi dikembangkan oleh Ramanathan pada tahun 2004. Dalam penelitian ini, mengubah kata-kata dengan menghapus *suffix* untuk pencarian informasi. Willet P mengusulkan algoritma *stemming* untuk perpustakaan elektronik dan sistem informasi pada tahun 2006. Zahurul MD mengembangkan *stemmer* untuk bahasa Bengali pada tahun 2009 untuk memeriksa ejaan bahasa Bengali. *Affix-exception* berbasis Urdu *stemmer* dikembangkan oleh Qurat-Ul-Ain Akram.

Tipologi morfologi bahasa untuk pencarian informasi ditemukan oleh Pirkola A. Dalam penelitian ini, IRS mengambil informasi berdasarkan tipologi morfologi. Pada tahun 1996, Hull D mengembangkan studi kasus algoritma *stemming* untuk evaluasi terperinci guna mengevaluasi kinerjanya.

Diusulkan oleh Mudassar untuk menemukan kata Marathi tersembunyi di *Knowledge Discovery Database* (KDD). Entitas yang diberi nama dalam bahasa Telugu menggunakan fitur yang bergantung pada bahasa dan pendekatan berbasis aturan dikembangkan oleh Sridhar B. Di tahun 2011, Model yang diusulkan menggunakan *Named Entity Recognition* (NER) untuk membendung kata-kata Telugu. Juhi Ameta memperkenalkan *stemmer* ringan untuk Gujarati di tahun 2012. Dalam model yang diusulkan ini, algoritma penanda akhiran digunakan untuk membendung kata-kata Tamil ke akar kata-katanya.

Algoritma Maulik digunakan untuk membendung kata-kata Hindi. Pada tahun 2012 sebuah *stemmer* iteratif untuk Bahasa Tamil diusulkan oleh Vivekanandan Ramachandran, dkk [3].

III. MORFOLOGI BAHASA TETUN

Bahasa Tetun merupakan bahasa resmi dan sudah tercantum di Konstitusi Negara Timor Leste, untuk digunakan sebagai bahasa kerja. Namun bahasa Tetun sendiri masih ada kata yang belum ditemukan oleh alih bahasa Tetun, sehingga masih ada kata yang meminjam dari bahasa lain seperti bahasa Portugis dan bahasa Indonesia. Timor Leste dipengaruhi oleh beberapa bahasa dikarenakan negara Timor Leste pernah dijajah oleh negara lain, seperti negara Portugal dan Indonesia, sehingga dalam bahasa Tetun sulit ditemukan kata-kata seperti awalan, akhiran, awalan dan akhiran, sisipan dan reduplikasi, dibandingkan dengan bahasa Indonesia.

Tetun yang digunakan dalam penelitian ini adalah Tetun campuran yaitu Tetun Frasa (Tetun Dili) dan Tetun Terik (Tetun Suai dan Viqueque), dikarenakan terdapat kata-kata yang tidak ada di Tetun Frasa dan begitu juga di Tetun Terik, seperti kata imbuhan awalan, imbuhan akhiran, imbuhan awalan dan akhiran, sisipan, dan kata reduplikasi.

Awalan yang diperbolehkan dalam pembentukan bahasa Tetun adalah sebagai berikut:

1. Aturan awalan (prefiks)
Aturan awalan yang diperbolehkan sebagai berikut:
   - Ha - + Tún = Hatún
     Contoh awalan kata dari **Ha-** sebagai berikut:
     Nita *tún* husi kareta = Nita turun dari mobil
     Nita *hatún* husi kareta = Nita diturunkan dari mobil
   - Na- + Kurut = Nakurut
     Contoh awalan kata dari **Na-** sebagai berikut:
     Hau hare hena ne,e *kurut* tiha ona = Saya lihat kain itu sudah kusut
     O nian hena *nakurut* tiha ona = Kain kamu sudah kusut
   - Nak- + Lees = Naklees
     Contoh awalan kata dari **Nak**- sebagai berikut:
     Hau *lees* tiha ona suratahan = Saya sudah merobek kertas
     Suratahan *naklees* tiha ona = kertas sudah dirobek
   - Nam- + kari = Namkari
     Contoh awalan kata dari **Nam**- sebagai berikut:
     Giterus ba *kari* aifunan iha Rate = Giterus pergi menabur bunga di makam
     Aifuna *namkari* deit = Menyebar bunga saja

   Aturan khusus untuk awalan yang diperbolehkan seperti **Na-, Nak-, Nam-,** contoh sebagai berikut:
   - **Na-** digunakan untuk barang yang sudah kusut tapi bisa diperbaiki
     Contoh: baju itu sudah kusut
   - **Nak-** digunakan untuk sesuatu yang tidak bisa diperbiki
     Contoh: Surat itu sudah disobek oleh adik
   - **Nam-** digunakan untuk sesuatu yang dilakukan masih tertunda
     Contoh: nanti kami akan menabur bunga di makam
   - Aturan awalan yang tidak diperbolehkan

Dór, N, Tén

2. Aturan akhiran (sufiks) yang diijinkan
   - Halimar + -dór = Halimardór
     Contoh akhiran kata -**Dór** adalah sebagai berikut:
     Nita ema nebe'e *halimardór* = Nita orang yang suka bercanda
   - Tunu + -n = Tunun
     Contoh akhiran kata -**N** adalah sebagai berikut:
     Hau **tunun** batar = Saya membakar jagung
   - Bosok + tén = Bosoktén
     Contoh akhiran kata -**Tén** adalah sebagai berikut:
     Mane ne'e b*osoktén* = Pria itu pembohong
   - Akhiran kata yang tidak diijinkan
     Ha, Na, Nak, Nam
3. Aturan awalan dan akhiran (konfiks)
   Gabungan kata awalan, tengah, dan akhiran untuk membentuk kata baru yang berhubungan dengan kata yang pertama
   - Ma + husu + k = Mahusuk
     Contoh kata awalan dan akhiran sebagai berikut:
     Hau sei *husu* Maria = Saya akan tanyakan Maria
     Maria sei *Mahusuk* Sira = Maria akan tanyakan mereka
   - Mak + sala + k = Maksalak
     Contoh kata awalan dan akhiran sebagai berikut:
     Nian *sala* tiha ona = Dia sudah salah
     Nian *Maksalak* hela = Dia orang bersalah
4. Aturan kalimat sisipan (infiks)
   Gabungan kata awal, tengah, dan akhir untuk membentuk kata baru yang beda arti dengan kata yang pertama. Contohnya sebagai berikut:
   - Babadók = Ba + dók = badók
     Contoh kalimat sisipan (infiks) sebagai berikut:
     Babadók = Badók dan sisipan **Ba**, contoh sebagai berikut:
     Hau sei *badók* husi fatin ne'e = Saya akan pergi jauh dari tempat ini
     Maria nia baku hela *Babadók* = Maria sedang menabuh drum
   - Aitahan = ta + han = aihan
     Contoh kalimat sisipan (infiks) sebagai berikut:
     Aitahan = Aihan dan sisipan **Ta**, contoh sebagai berikut:
     Labarik foti tiha *Aitahan* ne'e = Anak ambilkan daun ini
     Ita sempre persija *Aihan* = Kita selalu membutuhkan makanan
     Dalam kata bahasa Tetun untuk kata yang digunakan sebagai kata sisipan (infiks) hanya berapa kata saja.
5. Aturan reduplikasi
   Aturan reduplikasi ada dua yaitu reduplikasi sama arti dan reduplikasi beda arti sebagai berikut:
   a. Reduplikasi sama arti
      - Boot-boot
        Boot = besar
        Boot-boot = besar-besar
      - Funan-funan
        Funan = bunga
        Funan-funan = bunga-bunga
      - Barak-barak
        Barak = banyak
        Barak-barak = banyak-banyak
      - Idak-idak
        Idak = satu
        Idak-idak = satu-satu
      - Livru-livru
        Livru = buku
        Livru-livru = buku-buku

   b. Reduplikasi beda arti
      - Hotu-hotu
        Hotu = habis
        Hotu-hotu = semuanya
      - Fila-fila
        Fila = balik
        Fila-fila = berkali kali
      - Oin-oin
        Oin = muka
        Oin-oin = bermacam-macam
      - Ikus-ikus
        Ikus = belakang
        Ikus-ikus = akhir-akhir
      - Liu-liu
        Liu = lewat
        Liu-liu = diutamakan
6. Aturan kalimat pihak ketiga
   Kalimat ini digunakan untuk kata ke ketiga pihak atau juga bisa menunjukkan banyaknya barang.
   - Tauk
     Tauk = takut
     Hatauk = menakutkan saya
     Natauk = menakutkan kita
     Ratauk = menakutkan mereka
   - Dook
     Dook = jauh
     Hadook = jauhkan saya
     Nadook = jauhkan dari dia
     Radook = jauhkan dari mereka
   - Han
     Han = makan
     Hahan = saya makan
     Nahan = dia makan
     Rahan = mereka makan
   - Falun
     Falun = bungkus
     Hafalun = saya bungkus
     Nafalun = dia bungkus
     Rafalun = mereka bungkus
   - Toba
     Toba = tidur
     Hatoba = ditidurkan
     Natoba = dia tidur
     Ratoba = mereka tidur

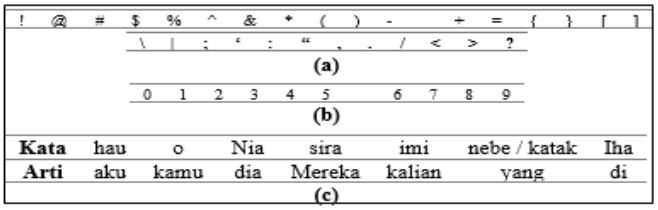
IV. DESAIN SISTEM

Bahan utama penelitian ini adalah kata-kata dari bahasa Tetun murni. Sistem ini membutuhkan daftar kata dasar yang dimasukkan dalam kamus kata. Kamus kata dasar berguna untuk mengetahui apakah kata yang dimasukkan dalam sistem sudah merupakan kata dasar atau belum dan untuk menentukan apakah hasil *stemming* berhasil atau tidak. Jumlah kata dasar yang dipakai dalam kamus kata adalah 176 kata. Kata dasar tersebut diperoleh dari pakar bahasa Tetun Timor Leste. Lalu kata-kata yang diberi imbuhan untuk diuji berjumlah 211.

Peneliti menekankan bahwa dalam hampir semua kata-kata bahasa Tetun yang terbentuk dari kata dasar dan imbuhan hanya mengandung kata dasar utuh. Tidak ada kata dasar yang mengalami perubahan bentuk saat diberi imbuhan.

Sebelum melalui tahap *stemming*, data teks harus melewati proses *pre-processing*. Tahap *pre-processing* terdiri dari tahap filterisasi dan tahap tokenisasi.

Tahap filterisasi bertujuan untuk menghilangkan tanda baca, angka, dan kata yang tidak penting (tidak bisa diberi awalan, akhiran, dan konfiks) seperti pada Gambar 1.



Gambar 1. Filterisasi : (a) Tanda Baca (b) Angka (c) Kata

Tahap pertama dalam proses *stemming* bahasa Tetun adalah mencari imbuhan awalan pada suatu kata hasil dari tokenisasi. Setiap kata akan dicari sub kata yang sesuai dengan daftar imbuhan awalan. Jika telah ditemukan maka akan dihapus dari kata tersebut untuk membentuk kata dasar. Daftar imbuhan awalan dalam bahasa Tetun seperti pada Tabel 1.

Tabel 1. Daftar Imbuhan Awalan

Imbuhan	Contoh Kata	Kata Dasar	Terjemahan
ha-	**ha**balun	balun	separuh
na-	**na**badak	badak	pendek
nak-	**nak**belit	belit	lengket
nam-	**nam**kaer	kaer	memegang

Tahap selanjutnya dalam proses *stemming* bahasa Tetun adalah mencari imbuhan akhiran pada suatu kata hasil dari tokenisasi. Setiap kata akan dicari sub kata yang sesuai dengan daftar imbuhan akhiran. Jika telah ditemukan maka akan dihapus dari kata tersebut untuk membentuk kata dasar. Daftar imbuhan akhiran dalam bahasa Tetun seperti pada Tabel 2.

Tabel 2. Daftar Imbuhan Akhiran

Imbuhan	Contoh Kata	Kata Dasar	Terjemahan
-dór	badinas**dór**	badinas	rajin
-n	tunu**n**	tunu	bakar
-tén	naok**tén**	naok	pencuri

Tahap ketiga dalam proses *stemming* bahasa Tetun adalah mencari imbuhan awalan sekaligus imbuhan akhiran pada suatu kata. Setiap kata akan dicari sub kata yang sesuai dengan daftar imbuhan awalan dan imbuhan akhiran. Jika telah ditemukan maka akan dihapus dari kata tersebut untuk membentuk kata dasar. Daftar imbuhan awalan dan akhiran dalam bahasa Tetun seperti pada Tabel 3.

Tabel 3. Daftar Imbuhan Awalan dan Akhiran

Imbuhan		Contoh Kata	Kata Dasar	Terjemahan
Awalan	Akhiran			
da-	-nuluk	**da**lima**nuluk**	lima	kelima puluh, lima
da-	-k	**da**tolu**k**	tolu	ketiga, tiga
ma-	-k	**ma**halo**k**	halo	membuatkan, buat
mak-	-k	**mak**sala**k**	sala	bersalah, salah

Tahap selanjutnya dalam proses *stemming* bahasa Tetun adalah mencari kata sisipan pada suatu kata. Setiap kata akan dicari sub kata yang sesuai dengan daftar imbuhan sisipan. Jika telah ditemukan maka akan dihapus dari kata tersebut untuk membentuk kata dasar. Tabel 4.

Tabel 4. Daftar Imbuhan Sisipan

Imbuhan	Contoh Kata	Kata Dasar	Terjemahan
-ba-	ba**ba**dók	badók	drum dan pergi jauh
-ta-	ai**ta**han	aihan	daun dan makanan
-k-	ha**k**maten	hamaten	berdiam dimatikan

Tahap selanjutnya dalam proses *stemming* bahasa Tetun adalah jika ada tanda hubung "-" di antara dua kata yang saling bertempelan tanpa spasi maka dianggap sebagai kata reduplikasi. Kata reduplikasi dalam bahasa Tetun tidak ada yang mengalami pergantian huruf vokal di salah satu dari dua kata reduplikasi. Contoh daftar kata reduplikasi bahasa Tetun seperti pada Tabel 5.

Tabel 5. Contoh Kata Reduplikasi

Reduplikasi	Kata Dasar	Terjemahan
barak-barak	barak	banyak-banyak
bikan-bikan	bikan	piring-piring
fila-fila	fila	berkali-kali, kembali
foun-foun	foun	baru-baru, baru

Algoritma *stemmer* untuk mencari kata dasar bahasa Tetun adalah sebegai berikut :

1. Tahap pertama memasukan kumpulan teks bahasa Tetun.
2. Tahap *filtering* untuk menghapus semua tanda baca dan huruf yang tidak penting.
3. Tahap tokenisasi untuk memisah teks menjadi bentuk kata-kata.
4. Proses *stemming*

a. Tahap pertama dalam proses *stemming* adalah memeriksa apakah variabel kata ada yang cocok dengan kata dasar dalam kamus. Jika variabel kata cocok dengan daftar kata dasar di kamus maka variabel kata ditampilkan tanpa melewati proses *stemming*.
b. Tahap kedua dalam proses *stemming* adalah memeriksa apakah ada konfiks (gabungan imbuhan awalan/prefiks dan akhiran/sufiks). Jika pada variabel kata terdeteksi adanya konfiks maka lanjut ke tahap 4.b.i. Namun jika tidak terdeteksi adanya konfiks maka lanjut ke tahap 4.c.
   i. Hapus imbuhan awalan (prefiks) "mak-", "da-", "ma-".
   ii. Lalu hapus imbuhan akhiran (sufiks) "-nuluk", "-k", dan
   iii. Menuju ke tahap 5.
c. Tahap ketiga adalah memeriksa apakah ada gabungan imbuhan awalan dan akhiran yang dilarang.
   i. Jika ada imbuhan awalan dan akhiran yang dilarang maka periksa apakah ada konfiks. Jika ada maka dilanjutkan ke tahap 4.b.i.
   ii. Namun jika konfiks tidak ditemukan maka periksa apakah ada gabungan imbuhan awalan dan akhiran yang dilarang.
      1. Jika ada gabungan imbuhan awalan dan akhiran yang dilarang maka periksa apakah ada imbuan akhiran (sufiks) "-dór" dengan disertai suku kata "ha-" di awal variabel kata.
         - Jika ada maka hapus imbuan akhiran (sufiks)"-dór". Lalu menuju ke tahap 5.
         - Jika tidak ada) maka hapus imbuhan awalan (prefiks) "nak-", "nam-", "ha-", "na-". Lalu menuju ke tahap 5.
      2. Jika tidak ada gabungan imbuhan awalan dan akhiran yang dilarang maka hapus imbuhan akhiran (sufiks) "-dór", "-n", "tén". Lalu menuju ke tahap 5.
d. Jika tidak ada imbuhan awalan dan akhiran yang dilarang maka periksa apakah ada sisipan (infiks).
   i. Jika ada sisipan (infiks) maka hapus sisipan tersebut. Lalu menuju ke tahap 5.
   ii. Jika tidak ada sisipan (infiks) maka hapus imbuhan awalan (prefiks) "nak-", "nam-", "ha-", "na-". Kemudian dilanjutkan dengan menghapus imbuhan akhiran (sufiks) "-dór", "-n", "tén". Lalu menuju ke tahap 5.

Tahap selanjutnya adalah memeriksa apakah variabel kata dasar hasil *stemming* ada yang cocok dengan kata dasar dalam kamus. Jika ada yang cocok, maka kata hasil *stemming* ditampilkan dengan status "cocok". Namun jika tidak ada yang cocok, maka hasil *stemming* ditampilkan dengan status "tidak ada di kamus". Uji coba yang dilakukan terdiri dari *stemming* imbuhan awalan, akhiran, konfiks, dan infiks. Setiap uji coba diberikan maksimal lima contoh kata. Isi pada kolom Hasil *Stemming* merupakan hasil *stemming* disertai dengan poin tahap proses *stemming* seperti algoritma *stemming* bahasa Tetun yang telah dijelaskan. Pengumpulan data diperoleh dari artikel, surat kabar, atau majalah yang bertemakan tentang pendidikan dalam bahasa Tetun. Penggunaan bahasa Tetun akan disesuaikan dengan kamus bahasa Tetun seperti Luís Costa Língua Tétum contributors para uma gramática. dan ahli yang akan menvalidasi penelitian ini adalah Lúis Costa [14].

1. Eliminasi Imbuhan Awalan
   Eliminasi awalan "ha-"

Tabel 6. Uji Coba Eliminasi Awalan "ha-"

Kata	Awalan	Hasil *Stemming*
habokon	ha	[Poin 4.c.i.1.b] => bokon (Cocok)
habaruk	ha	[Poin 4.d.ii] => baruk (Cocok)
habeik	ha	[Poin 4.d.ii] => beik (Cocok)
habelit	ha	[Poin 4.d.ii] => belit (Cocok)
hafaluk	ha	[Poin 4.d.ii] => faluk (Cocok)

2. Eliminasi Imbuhan Akhiran
   Eliminasi akhiran "-dór"

Tabel 7. Uji Coba Eliminasi Akhiran "-dór"

Kata	Awalan	Hasil *Stemming*
Hemudór	Dór	[Poin 4.d.ii] => badinas (Cocok)
Devedór	Dór	[Poin 4.d.ii] => deve (Cocok)
Fumadór	Dór	[Poin 4.d.ii] => fuma (Cocok)
Hamnasadór	Dór	[Poin 4.c.i.1.a] => hamnasa (Cocok)
Pescadór	dór	[Poin 4.d.ii] => pesca (Cocok)

3. Eliminasi infiks "-ka-"

Tabel 8. Uji Coba Eliminasi Infiks "-k-"

Kata	Awalan	Hasil *Stemming*
Hakmate	K	[Poin 4.d.i] => hamate (Cocok)

Contoh proses *stemming* "habaruk" dengan awalan "ha-":

a. Periksa kata "habaru" apakah cocok dengan kata pada kamus. Hasilnya adalah tidak ada. Maka lanjut ke tahap b.
b. Periksa kata "habaruk" apakah ada konfiks (gabungan imbuhan awalan (prefiks) dan akhiran (sufiks)) berdasarkan data pada Tabel 3.3. Hasilnya adalah tidak ada. Maka lanjut ke tahap c.
c. Tahap ketiga adalah memeriksa apakah ada gabungan imbuhan awalan (prefiks) dan akhiran (sufiks) yang dilarang. Hasilnya adalah tidak ada. Maka periksa apakah ada sisipan (infiks).
   i. Hasilnya adalah tidak ada sisipan (infiks) maka hapus imbuhan awalan (prefiks) "nak-", "nam-", "ha-", "na-". Kemudian dilanjutkan dengan menghapus imbuhan akhiran (sufiks) "-dór", "-n", "tén". Hasilnya adalah "baruk"
   ii. Lalu menuju ke tahap d.

d. Periksa apakah daftar kata dasar pada kamus ada yang cocok dengan kata "baruk". Hasilnya adalah cocok. Maka tampilkan kata "baruk", status cocok, dan poin "4.d.ii"

Contoh proses *stemming* "dalimanuluk" dengan awalan "da-"dan akhiran "-nuluk":

a. Periksa kata "dalimanuluk" apakah cocok dengan kata pada kamus. Hasilnya adalah tidak ada. Maka lanjut ke tahap b.
b. Periksa kata "dalimanuluk" apakah ada konfiks (gabungan imbuhan awalan/prefiks dan akhiran/sufiks) berdasarkan data pada Tabel 3.3. Hasilnya adalah ada. Maka lakukan proses berikut :
    i. Hapus imbuhan awalan (prefiks) "mak-", "da-", "ma-". Hasilnya adalah "limanuluk".
    ii. Lalu hapus imbuhan akhiran (sufiks) "-nuluk", "-k". Hasilnya adalah "lima". (Poin 4.b.ii)
    iii. Menuju ke tahap c.
c. Periksa apakah daftar kata dasar pada kamus ada yang cocok dengan kata"lima". Hasilnya adalah cocok. Maka tampilkan kata "lima", status cocok, dan poin "4.b.ii"

Uji coba dilakukan dengan data kamus kata dasar berjumlah 176. Sedangkan jumlah kata yang diuji berjumlah 211. Perhitungan akurasi menggunakan persamaan (1).

$$Akurasi = \left(\frac{\sum KD}{\sum K}\right) \times 100\% \quad (1)$$

Dimana $\sum KD$ adalah jumlah kata dasar yang berhasil melalui proses *stemming*, dan $\sum K$ adalah jumlah kata yang melalui uji *stemming*. Akurasi dinyatakan dalam bentuk persentase (%). Jumlah *stemming* benar/cocok 191, jumlah *stemming* salah/tidak cocok 20, akurasi 90.52 %, dan rata-rata waktu *stemming* semua kata adalah 0.0182 milidetik. Berdasarkan perhitungan menggunakan persamaan (1) maka didapat akurasi sebesar 90.52%. Penyebab kegagalan pada uji coba adalah karena ada kasus *spelling exception* dan kasus *overstemming*. *Spelling exception* adalah kasus dimana kata dasar masih memiliki imbuhan setelah melewati proses *stemming*. Sedangkan *overstemming* adalah kasus dimana kata dasar mengalami pengurangan huruf atau suku kata yang dianggap sebagai imbuhan akibat dari proses *stemming*.

Contoh kasus *spelling exception* seperti proses *stemming* awalan "ha-" pada kata "hatanis" menjadi "hanis" padahal yang benar adalah "tanis". Awalan "ha-" masih tetap ada, sedangkan ada huruf atau suku kata (bagian kata dasar) yang terhapus. Penyebab kegagalan adalah karena ada bagian huruf yang dianggap sebagai sisipan ("-ta-") yang dihapus, sedangkan awalan "ha-" diabaikan oleh algoritma.

Contoh kasus *overstemming* seperti proses *stemming* awalan "na-" pada kata "nafatin" menjadi "fati" padahal yang benar adalah "fatin". Yaitu ada awalan "na-" sudah berhasil terhapus bersama kasus terhapusnya sebagian huruf dari kata dasar. Penyebab kegagalan adalah karena ada bagian huruf yang dianggap sebagai akhiran "-n" yang dihapus, dan ada bagian huruf yang dianggap sebagai awalan.

V. KESIMPULAN

Berdasarkan uraian penjelasan di atas, maka dapat diambil kesimpulan bahwa algoritma *stemming* bahasa Tetun memiliki tahap utama memeriksa konfiks, gabungan awalan & akhiran yang dilarang, dan sisipan secara urut. Kegagalan algoritma *stemming* bahasa Tetun disebabkan karena kesalahan *overstemming*. Algoritma *stemming* bahasa Tetun mendapatkan akurasi sebesar 90.52%.

REFERENSI

[1] Costa, L. (2015). *Língua Tétum - contributos para uma gramática, Edições Colibri*. Lisboa.
[2] Adriani, M., Asian, J. & Nazief, B. (2007). Stemming Indonesian: A Confi x-Stripping Approach. *ACM Transactions on Asian Language Information Processing*, Vol. 6, No. 4, Article 13. 1-13.33.
[3] Thangarasu, M & Manavalan, R (2013). *A Literature Review: Stemming Algorithms for Indian Languages*. Jurusan Ilmu Komputer dan Aplikasi Universitas Seni Rupa dan Sains KSRangasamy.