Jurnal Linguistik Komputasional
Indonesian Association for Computational Linguistics
JLK
Vol 2, No 1
Maret 2019

Sistem Identifikasi Bahasa Jawa dan Bahasa Indonesia Dokumen Teks Berbasis N-Gram Karakter

Fidelia Vera Sentosa[#1], Lucia Dwi Krisnawati[#2], Aditya Wikan Mahastama[#3]

[3]*Informatika, Universitas Kristen Duta Wacana*
Jl. Dr. Wahidin Sudiro Husodo No. 5-25, Yogyakarta
[1]fideliavera@ti.ukdw.ac.id
[2]krisna@staff.ukdw.ac.id
[3]mahas@staff.ukdw.ac.id

Abstrak— **Identifikasi bahasa adalah sebuah proses yang mencoba menemukan bahasa yang digunakan dalam sebuah wacana secara otomatis. Sistem Identifikasi Bahasa (SIB) pada dasarnya dibedakan menjadi SIB lisan yang mengidentifikasi bahasa tutur lewat fitur akustik atau fonem, dan SIB berbasis fitur grafem dalam berbagai level dan kategori linguistiknya. Penelitian ini mencoba untuk membangun SIB yang dirancang untuk membedakan teks berbahasa Jawa dari bahasa Indonesia dan bahasa lainnya. Profil bahasa yang digunakan dibangun dari korpus yang diambil dari korpus Trawaca dan beberapa sumber daring dari berbagai topik. Tujuannya adalah untuk memperkaya kosa kata dan meningkatkan jumlah tipe kata. Profil bahasa tiap kategori dibentuk dari n-gram berbasis karakter dan diambil 100 n-gram dengan nilai CF tertinggi. Perhitungan jarak antara profil bahasa dengan dokumen uji dilakukan dengan menggunakan ukuran Out-Of-Place (OOP). Hasil pengujian menunjukkan bahwa Presisi idenfikasi dokumen berbahasa Jawa mencapai 0.96, sedangkan Presisi dokumen berbahasa Indonesia mencapai 0.86. Nilai Akurasi total identifikasi mencapai 0.85. Nilai Presisi identifikasi bahasa Indonesia jauh lebih rendah dibanding nilai Presisi identifikasi bahasa Jawa disebabkan diujikannya dokumen berbahasa Melayu-Malaysia yang tentu saja teridentifikasi sebagai dokumen berbahasa Indonesia. Waktu rata-rata identifikasi untuk dokumen pendek maupun panjang mencapai 3 detik. Penelitian ini berhasil mendemonstrasikan bahwa dengan fitur yang minimal, maka presisi identifikasi yang tinggi bisa tercapai dengan waktu proses yang relatif pendek dan stabil.**

Kata kunci*— *Identifikasi Bahasa, N-gram karakter, profil bahasa Jawa, out-of-place ranking

I. PENDAHULUAN

Identifikasi bahasa merupakan suatu proses yang mencoba menemukan bahasa yang digunakan dalam sebuah wacana secara otomatis. Wacana tersebut bisa berbentuk teks tertulis maupun kumpulan percakapan lisan [1] yang dikenal juga sebagai korpus tutur. Identifikasi bahasa secara otomatis menjadi salah satu tahap pra-pemrosesan dalam aplikasi-aplikasi Pengolahan Bahasa Alami (NLP) seperti dalam Terjemahan Mesin (Machine Translation) [2] , pembersihan data dari bahasa asing [3], klasifikasi dokumen [4], maupun dalam panggilan telefon gawat darurat [2].

Sekalipun Baldwin dan Lui [5] mencatat bahwa riset Identifikasi Bahasa otomatis telah dimulai sejak tahun 1967, namun topik ini tetap dikembangkan karena adanya kebutuhan terhadap aplikasi identifikasi bahasa serta masih adanya ketidak-puasan akan kemampuan sistem untuk mengenali bahasa dalam dokumen pendek maupun dokumen multi-bahasa [3]. Majliš [1] mencatat bahwa jumlah bahasa dalam multi-dokumen yang dikenali berbanding terbalik dengan tingkat akurasi pengenalan sistem. Demikian juga dengan panjang dokumen, dimana semakin pendek dokumen, akurasi pengenalanpun akan semakin menurun.

Persoalan lain yang dihadapi oleh sistem identifikasi bahasa adalah kecepatan proses yang mutlak diperlukan dalam pengenalan langsung (*real time*). Kecepatan proses ini ditentukan oleh dua faktor, yakni jumlah profil yang digunakan untuk merepresentasikan sebuah bahasa serta algoritma yang diterapkan.

Untuk itulah, penelitian ini berusaha menggunakan profil bahasa seminimal mungkin, namun profil tersebut diharapkan mampu memberikan tingkat akurasi yang tinggi bahkan bagi dokumen pendek. Sistem yang dibangun dirancang untuk mengenali dwi bahasa, yakni bahasa Jawa dan bahasa Indonesia. Teks berbahasa lain akan dikategorikan sebagai kelas ‘bahasa lainnya’. Jarak antara profil sebuah bahasa dengan bahasa teks yang diujikan dihitung berdasarkan perhitungan perangkinan *out of place* yang diperkenalkan olek Cavnar dan Trenkle di [4].

Berbeda dari Cavnar dan Trenkle [4], penelitian ini menggunakan n-gram dengan n = 2-5. Tiap kelas bahasa diwakili oleh 100 profil bahasa gabungan dari bigram, trigram, quadgram dan pentagram tertinggi. Dengan

Submitted 10-02-2019; Revised 20-03-2019; Accepted 25-03-2019

demikian jumlah profil ini jauh lebih kecil dari jumlah profil yang digunakan di [4]. Profil bahasa tiap kelas dihasilkan dari masing-masing 100 n-gram dengan frekuensi tertinggi, sehingga jumlah awal calon profil tiap bahasa ada 400 n-gram. Namun dari ke-400 n-gram ini disaring untuk mendapatkan hanya 100 profil yang tertinggi. Dengan demikian, kontribusi penelitian ini terletak pada jumlah minimal profil yang digunakan untuk proses identifikasi dengan waktu identifikasi yang konstan, relatif pendek, dan yang tidak dipengaruhi oleh panjang dokumen.

II. TINJAUAN PUSTAKA

Berdasarkan survei yang kami lakukan, Sistem Identifikasi Bahasa (SIB) bisa dibedakan berdasarkan fitur atau profil yang digunakan serta metode yang diterapkan dalam pengenalan bahasa sebuah wacana. Fitur pengenalan yang digunakan pada dasarnya bisa dikategorikan menjadi fitur grafem, yang banyak digunakan oleh SIB untuk dokumen teks, dan fitur akustik (fonem) yang dibutuhkan oleh SIB dengan masukan berbentuk ujaran percakapan. Beberapa SIB lisan bergantung pada fitur fonem dan prosodi [6], fitur prosodi dan forman yang merupakan puncak-puncak atau nilai frekuensi maksimal dalam gelombang suara [7]. Dehak, dkk [8] memperkenalkan penggunaan fitur i-vector bagi SIB lisan yang menjadi fitur termutakhir sampai kini. I-vector didapatkan dengan cara mensegmentasi frekuensi bunyi tiap 20 milidetik dalam sebuah jendela. Bunyi di tiap jendela tersebut akan disegmen kembali dalam 10 milidetik. Tiap fitur ini kemudian dikonversikan menjadi vector dan dinormalkan melalui standard distribusi normal. Sekalipun i-vector banyak digunakan untuk pengenalan pembicara, namun Dehak et al. [8] dan Dominguez et al. [2] menggunakannya untuk mendeteksi bahasa dalam SIB lisan.

Berbeda dari SIB lisan, SIB dokumen teks menekankan penggunaan grafem di tingkat yang berbeda. Satuan grafem terkecil yang dikenal sebagai hurufpun bisa digunakan sebagai fitur SIB [1] [9], sedangkan perpaduan antara huruf dan diakritik dijadikan fitur minimal oleh Takçi et al. [10]. Penggunaan n-gram berbasis karakter, yang merupakan rangkaian huruf sejumlah n yang berurutan, diperkenalkan oleh Cavnar dan Trenkle [4]. Sedangkan fitur SIB di level kata digunakan oleh Selamat dan Akosu [11] serta Sarma et al. [12], fitur kata dalam kalimat diperkenalkan di [12] [13]. Baldwin dan Lui [5] memperkenalkan penggunaan *byte n-gram* dan *codepoint n-gram*, di mana *codepoint n-gram* diperoleh dari konversi karakter ke kode utf-8 dan unicode [5]. Byte n-gram juga diterapkan di penelitian Liu et al [3].

Metode yang digunakan dalam SIB bisa diklasifikasikan dalam pengunaan statistik murni dan Pembelajaran Mesin (*Machine Learning*). Pendekatan statistik murni biasanya menggunakan frekuensi kemunculan fitur dan perankingan. Salah satunya adalah dengan menghitung jumlah kemuculan kata yang terdapat di tiap leksikon bahasa latih. Jumlah kata yang terkandung dalam leksikon bahasa tertinggi itu dijadikan ukuran menentukan bahasa sebuah teks [11]. Percepatan komputasi dilakukan dengan melakukan segmentasi leksikon tiap bahasa berdasarkan panjang kata, sehingga perbandingan dilakukan hanya pada segmen leksikon dengan dengan panjang kata yang sama [11]. Berbeda dari [11], Takcı dan Soğukpınar [9] menghitung frekuensi serta distribusi huruf dalam sebuah teks, dan nilai teks atau bahasa dihasilkan dari komputasi nilai frekuensi, distribusi huruf serta pembobotan tiap huruf di dokumen dalam korpusnya. Metode serupa yang berbasiskan kamus diterapkan juga di [1].

Cavnar dan Trenkle [4] menggunakan frekuensi n-gram karakter dari keseluruhan korpus (collection frequency) untuk memilih fitur yang dijadikan profil dari tiap bahasa. Dia menggunakan 300 fitur dengan nilai CF tertinggi untuk dibandingkan dengan fitur dokumen teks. Jarak antara profil bahasa dengan dokumen uji dilakukan dengan penghitungan jarak ranking tiap profilnya dengan rumus *Out-of-Place* [4].

Pendekatan statistik yang dipadukan dengan semantik ditawarkan oleh Sarma et al [12]. Dalam penelitiannya mereka menggunakan Word2Vec melalui pustaka Glove untuk menkomputasi persamaan global dan lokal (global & local similarity) tiap kata.

Dalam survey kami, pendekatan pembelajaran mesin lebih dominan penerapannya dalam SIB lisan, sedangkan SIB dokumen teks masih menerapkan pendekatan pembelajaran mesin klasik seperti Nearest-Neighbour and Nearest-Prototype Models [5], Naive Bayes [5] [12], Support Vector Machines [1] [5] [10] [12], dan Linear Discriminant Analysis [10]. Penggunaan Jaringan Syaraf Tiruan seperti Recurrent Neural Network juga Deep neural Network diterapkan untuk SID lisan di [5] , namun Dehak et al [8] menerapkan LDA dan SVM beserta variannya.

III. PEMBENTUKAN PROFIL BAHASA

Identifikasi bahasa dianggap sebagai sebuah proses tertutup, artinya dengan diberi data dari tiap bahasa yang telah didefinisikan sebelumnya, sebuah sistem diharapkan mampu mengklasifikasikan bahasa dari teks yang diujikan [5]. Definisi di atas mensyaratkan bahwa sebuah SIB selalu membutuhkan data. Jika proses identifikasi ini dipandang sebagai proses klasifikasi, maka data tersebut digunakan untuk mengekstraksi fitur sebuah bahasa untuk membentuk sebuah kelas bahasa tertentu. Jika proses identifikasi dipandang sebagai proses pencocokan dan penghitungan jarak kemiripan, maka data terkumpul akan diekstrasi untuk membentuk profil dari tiap bahasa yang akan diidentifikasikan. Pada dasarnya profil dan fitur

bahasa memiliki fungsi yang sama yakni menjadi ciri dan parameter perbandingan tiap kelas atau kategori bahasa. Dalam sub-bab ini akan diuraikan pertama-tama tentang pengumpulan data, tahap pra-pemrosesan, n-gram serta pembentukan profil bahasa.

A. Pengumpulan Data

Penelitian ini mencoba membangun sistem identifikasi bahasa yang mampu mengidentifikasi 3 kelas bahasa yakni Bahasa Jawa, Bahasa Indonesia, serta bahasa lainnya. Bahasa lainnya ini disediakan sebagai kelas tampungan bagi setiap bahasa yang tidak termasuk dalam kedua bahasa yang disebutkan di awal.

Berdasarkan cakupan ini, maka data yang dipergunakan untuk membangun profil bahasa didapatkan dari dokumen-dokumen yang dikumpulkan secara manual. Beberapa cara yang dilakukan adalah:

- **Mengkopi sebagian data dari korpus Trawaca**. Korpus TRAWACA [14] berisi sekumpulan dokumen yang berbahasa Jawa dan Indonesia yang didapatkan dengan cara melakukan *Web Scrapping* otomatis pada laman Wikipedia Jawa, Wikipedia Indonesia, Sastra Lestari (sastra.org), dan Ki Demang.

  Menyalin artikel dari situs tertentu. Penyalinan artikel secara manual juga dilakukan dari beberapa situs sepert CNN Indonesia, Detik.com, cerita rakyat dan Puisi Indonesia. Adapun tujuannya adalah mendapatkan keragaman kosa kata dari berbagai ragam teks seperti berita di bidang politik, olah raga serta sastra dan budaya. Penambahan artikel ini diharapkan meningkatkan jumlah tipe kata yang nantinya memperkaya profil setiap bahas.

Data statistik asal dokumen yang didapatkan bisa dilihat di Tabel 1. Untuk dokumen yang diambil dari korpus Trawaca ini merupakan teks berbahasa Jawa dan ditulis dengan aksara Latin. Jumlah total dokumen mencapai 40 dan terdiri dari kurang lebih 35.600 token atau sekitar 125 KB (Kilo Byte) dalam ukuran berkas. Untuk sampel bahasa Indonesia yang bersumber dari korpus Trawaca, dipilih dokumen yang berasal dari wikipedia.org sebanyak 41 dokumen dengan berbagai tema, ditambah dengan 7 dokumen yang terambil dari halaman detik.com, 7 dokumen dari halaman cnnindonesia.com, 1 dokumen yang berisi 10 cerita rakyat Indonesia, dan 1 dokumen yang berisi 11 puisi Indonesia. Sehingga gabungan dari 57 dokumen berbahasa Indonesia tersebut menghasilkan kurang lebih 35.200 kata atau 196 KB (Kilo Byte) dalam bentuk file.

TABEL I
DATA STATISTIK BAGI SUMBER PENGUMPULAN DATA

Sumber data	Jumlah dokumen	Bahasa
Trawaca:		
Wikipedia Jv	38	Jawa
Sastra Lestari	8	Jawa
Wikipedia	41	Indonesia
Detik.com	7	Indonesia
Cnnindonesia.com	7	Indonesia
Cerita Rakyat Indonesia	1 (terdiri dari 10 cerita)	Indonesia
Puisi Indonesia	1 (11 puisi)	Indonesia

B. Tahap Pra-proses

Untuk mendapatkan profil dari kedua bahasa tersebut, maka dokumen dikelompokkan berdasarkan bahasanya. Kemudian, masing-masing kumpulan dokumen tersebut dilakukan normalisasi sebagai tahan pra-proses dengan cara:

- **Konversi huruf** yakni dengan mengubah semua huruf menjadi huruf kecil.
- **Penghapusan tanda baca dan angka,** sebagai akibatnya jika ada sebuah token yang terdiri dari huruf dan angka, maka angka dari token tersebut akan hilang dan tinggal string hurufnya saja, contoh ‘ke-2’ akan terkonversikan menjadi ‘ke’.
- **Konversi multi-spasi ke spasi tunggal**. Pengubahan ini dimaksudkan untuk mengubah ‘tab’, garis baru dan multi-spasi lainnya menjadi satu spasi tunggal saja.
- **Konversi spasi tunggal menjadi tanda garis bawah**, teknik ini dilakukan sehubungan dengan profil yang akan dibangun adalah n-gram berbasis karakter. Garis bawah dirasa menjadi penanda awal dan akhir sebuah token yang jauh lebih baik dan jelas dibandingkan dengan spasi kosong, atau tanpa spasi dalam pembentukan n-gram karakternya. Dengan demikian, garis bawah akan menandai bahwa n-gram tertentu terletak di awal atau akhir token.

Setelah tiap dokumen tersebut telah dinormalisasi, maka profil bahasa dalam bentuk n-gram siap dibentuk.

C. N-gram

N-gram adalah potongan karakter sepanjang n dari sebuah string yang lebih panjang [4]. Cara pemotongannya berurutan dari karakter pertama, kedua dan seterusnya hingga membentuk potongan yang saling bertumpang tindih. Pada dasarnya ada dua jenis n-gram, yaitu n-gram berbasis karakter dan n-gram kata. Jika n=1, maka hasilnya akan disebut sebagai unigram. Bigram mewakili gram dengan n=2, trigram untuk menamai string yang terbentuk dari n=3, dst. Sebagai contohnya, kita mempunyai sebuah string “BELAJAR_BACA”. String ini akan kita ubah menjadi unigram sampai pentagram karena n= 1-5. Dengan tujuan mempermudah pembacaan, maka dalam contoh ini tidak diberlakukan proses konversi huruf

besar ke huruf kecil, sehngga didapatkan n-gram sebagai berikut.

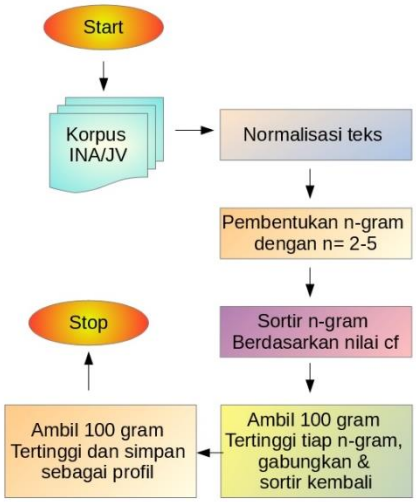
- Unigram = B, E, L, A, J, A, R, _, B, A, C, A
- Bigram= BE, EL, LA, AJ, JA, AR, R_, _B, BA, AC, CA
- Trigram= BEL, ELA, LAJ, AJA, JAR, AR_, R_B, _BA, BAC, ACA
- Quadgram: BELA, ELAJ, LAJA, AJAR, JAR_, AR_B, R_BA, _BAC, BACA
- Pentagram: BELAJ, ELAJA, LAJAR, AJAR_, JAR_B, AR_BA, R_BAC, _BACA

Dengan pembentukan n-gram karakter seperti yang dicontohkan di atas, maka dengan teks sepanjang m, kita akan mendapatkan n-gram sejumlah (m-n)+1, dengan catatan spasi diperhitungkan sebagai karakter kosong. Pembentukan n-gram berbasis kata tidak jauh berbeda dari cara pembentukan n-gram berbasis karakter. Hanya saja dibutuhkan sebuah pra-proses tambahan yakni tokenisasi yang memisahkan serangkaian karakter berdasarkan spasi kosong. Serangkaian karakter yang membentuk satu string inilah yang disebut token atau kata.

D. Pembentukan Profil Bahasa Indonesia dan Jawa

Untuk pembuatan profil n-gram karakter, langkah pertama adalah melakaukan tahap pra-proses seperti yang dijelaskan di sub-bahasan IIIB. N-gram yang dibentuk menjadi profil bahasa Jawa dan Indonesia adalah bigram, trigram, quadgram dan pentagram. Unigram tidak dipilih sebagai profil dengan asumsi bahwa distribusi karakter tunggal belum bisa menjadi profil yang kuat bagi bahasa yang sangat dekat seperti bahasa Indonesia dan Jawa.

Setelah konversi spasi menjadi tanda garis bawah (_), maka secara berurutan dilakukan pemotongan karakter sepanjang n= {2, 3, 4, 5} yang dimulai dari indeks ke-0. Iterasi dilakukan dengan menambahkan 1 ke indeksnya dan diakhiri setelah nilai indeks mencapai panjang teks dikurangi n di n-gramnya. Substring sebagai n-gram kemudian ditampung dalam tabel hash sementara, dengan string sebagai kuncinya. Bagi setiap n-gram yang sudah ada dalam tabel hash sebagai kunci, maka nilai frekuensinya akan ditambahkan satu, sehingga akan menghasilkan frekuensi kolektif (CF/count). Jika String n-gram tersebut belum terdapat dalam tabel hash, maka string tersebut ditambahkan sebagai kunci dan diberi nilai frekuensi 1. Setiap n-gram ditampung dalam tabel hash yang berbeda sehingga ada tabel tampung hash bigram, trigram, quadgram dan pentagram. Gambar 1 menunjukkan diagram alir pembentukan profil bahasa yang diterapkan dalam kumpulan dokumen berbahasa Indonesia dan Jawa.



Gambar 1 Diagram alir pembentukan profil masing-masing bahasa

Untuk masing-masing n-gram dilakukan penyortiran berdasarkan nilai CF-nya dan diurutkan dari gram dengan nilai CF tertinggi sampai nilai terendah. Dari masing-masing n-gram tersebut kemudian diambilah 100 gram tertinggi sehingga kita dapatkan 400 n-gram tertinggi. Ke-400 gram ini digabungkan menjadi satu, dan diurutkan kembali berdasarkan nilai CF, dan diambil 100 gram tertinggi. 100 String tertinggi beserta posisi rangkingnya kemudian disimpan untuk dijadikan profil bahasa. Dengan demikian profil bahasa terdiri dari campuran bigram, trigram, quadgram serta pentagram. Profil bahasa yang digunakan berjumlah 100 saja dengan tujuan mengurangi waktu komputasi perbandingan sehingga dimungkinkan pengecakan bahasa secara cepat.

IV. PROSES IDENTIFIKASI BAHASA

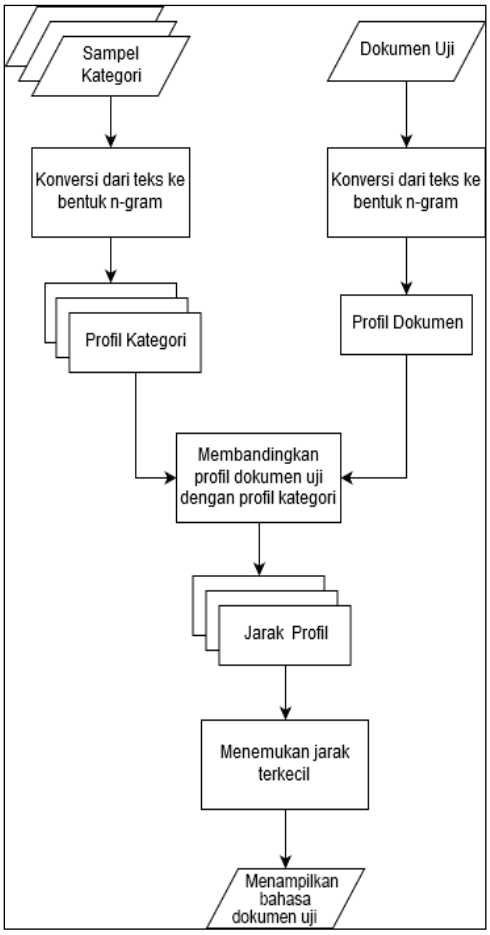
Proses selanjutnya adalah menggunakan profil bahasa tersebut untuk menguji sebuah teks apakah bahasa yang tertulis di dalamnya tersebut termasuk dalam kelas bahasa Indonesia, Jawa atau lainnya. Maka dibutuhkan sebuah metriks yang akan menghitung jarak dan dasar penghitungan jarak tersebut. Penelitian ini menggunakan metrik yang diusulkan oleh Cavnar dan Trenkle [4] yang cocok untuk penghitungan jarak profil n-gram berbasis karakter dan yang dinamai perangkingan jarak *out of place*.

Namun sebelumnya, proses tokenisasi, pembentukan n-gram, dan penyaringan n-gram menjadi profil dilakukan juga terhadap dokumen yang akan diidentifkasi bahasanya, dalam hal ini disebut sebagai dokumen uji. Dengan demikian proses identifikasi bahasa dijelaskan di Gambar 2.

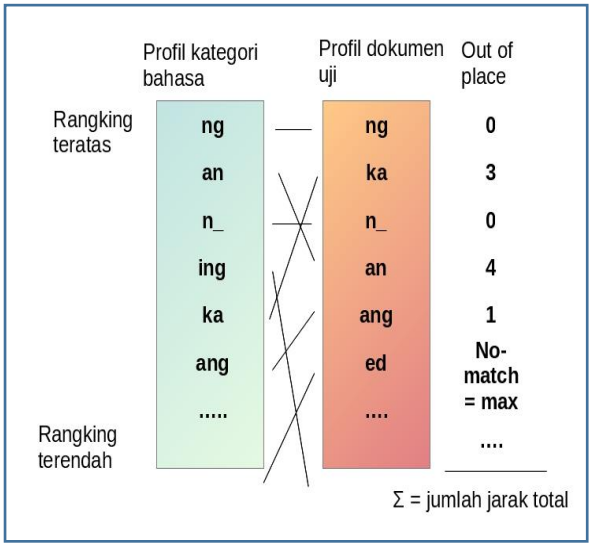
Proses perhitungan nilai *out-of-place* dapat dijelaskan sebagai berikut: Setelah menghitung jarak n-gram profil dokumen uji dengan kedua profil bahasa Jawa dan Indonesia, maka langkah selanjutnya adalah menentukan jarak profil terdekat dari kedua profil bahasa yang ada.

Pada setiap n-gram yang terdapat pada profil dokumen uji, dilakukan pencocokan untuk mencari n-gram yang sama pada sebuah profil bahasa dan menghitung seberapa jauh perbedaan tempat yang direpresentasikan sebagai jarak. Jarak antar profil dihitung dengan cara menjumlahkan setiap nilai *out of place*, seperti pada Gambar 3 berikut.

Ilustrasi di Gambar 3 menunjukkan penghitungan nilai *out of place* untuk setiap n-gram yang terdapat pada profil dokumen uji dan profil bahasa. Perlu diperhatikan bahwa posisi dalam Gambar 3 di atas diberikan sebagai contoh dan tidak menggambarkan kondisi ranking profil yang sesungguhnya. Sebagai contoh, n-gram 'ka' pada Gambar 3 terletak pada urutan 2 di profil dokumen uji dan terletak pada urutan 5 pada profil kategori bahasa. Nilai out-of-place dihitung dengan mengurangkan selisih ranking pada profil kategori bahasa dengan profil dokumen uji, yang kemudian diambil nilai absolutnya. Dengan demikian, nilai *out of place* dari n-gram 'ka' dalam contoh ini adalah 5-2 = 3. Jika terdapat n-gram dokumen uji yang tidak ada di profil kategori bahasa, maka nilai *out of place* diambilkan dari nilai maksimum, yaitu 100 – ranking terendah dalam profil kategori bahasa.



Gambar 2 Diagram alir proses identifikasi bahasa



Gambar. 3. Penghitungan jarak out of place, diadaptasi dari [4].

Jarak antara profil dokumen uji dengan sebuah profil kategori bahasa kemudian didapatkan dengan cara menjumlahkan nilai *out-of-place* (OOP) setiap n-gram pada kedua profil tersebut. Secara matematis, OOP yang merupakan metrik berbasiskan jarak ranking dihitung berdasarkan persamaan 1 [5] seperti berikut ini:

$$OOP(D_x, D_y) = \sum_{t \in D_x \cup D_y} ABS(R_{D_x}(t), - R_{D_y}) \quad (1)$$

dimana R_{Dx} (t) merupakan rangking gram *t* di dokumen Dx yang merupakan profil dokumen uji, sedangkan R_{Dy} (t) merujuk pada ranking gram yang sama di profil kategori bahasa.

Nilai OOP sebuah dokumen uji akan dihitung terhadap setiap profil kategori bahasa, yang pada kasus ini hanya terdiri dari profil bahasa Jawa dan bahasa Indonesia. Setelah mendapatkan kedua nilai OOP tersebut, langkah selanjutnya adalah menghitung selisih persentase bahasa dokumen uji terhadap kedua kategori tersebut, kemudian kedua persentase tersebut dibandingkan untuk mengetahui persentase manakah yang lebih kecil. Persentase yang lebih kecil, adalah hasil dari identifikasi bahasa dokumen tersebut. Akan tetapi, pada sistem identifikasi bahasa ini diberikan nilai ambang (*threshold)* sebesar 5%, jadi jika selisih persentase antara bahasa Jawa dan bahasa Indonesia kurang dari 5%, maka sistem akan menampilkan luaran bahwa bahasa dokumen uji termasuk dalam kategori bahasa lainnya. Apabila selisih persentase nilai OOP antara dokumen uji dengan kedua kategori tersebut melebihi 5%, maka sistem akan menyatakan bahwa bahasa dokumen adalah bahasa Jawa, jika nilai OOP dokumen uji terehadap profil kategori bahasa Jawa lebih kecil daripada nilai OOP dengan kategori bahasa Indonesia. Untuk penghitungan persentase selisih bahasa dituliskan dalam Persamaan (2).

$$persentase\ selisih = \frac{|OOP(D_x, D_{JV}) - OOP(D_x, D_{indo})|}{OOP(D_x, D_{JV}) + OOP(D_x, D_{indo})}\ X\ 100 \quad (2)$$

Sebagai contoh perhitungan bagi persamaan (2), nilai OOP(D_x, D_{JV}) atau sebuah dokumen uji D_x terhadap profil kategori bahasa Jawa (D_{JV}) mencapai 125, sedangkan nilai OOP(D_x, D_{Indo}) mencapai 225, maka prosentase selisih kedua nilai OOP tersebut adalah abs(125-225)/125+225 * 100. Hasilnya adalah 28.6%. Karena selisih prosentase diatas 5% dan nilai OOP ke profil kategori bahasa Jawa lebih rendah dari profil bahasa Indonesia, maka bahasa dokumen x diidentifikasi sebagai bahasa Jawa.

V. IMPLEMENTASI, PENGUJIAN DAN EVALUASI

A. *Implementasi sistem*

Untuk melakukan pengujian terhadap pendekatan yang diusulkan, dikembangkanlah sebuah sistem identifikasi Bahasa Jawa dan Bahasa Indonesia berbasis n-gram. Sistem ini dikembangkan menggunakan bahasa pemrograman PHP dan basis data MySQL. Sistem ini dapat menyimpan sampel-sampel dokumen penyusun profil bahasa (Jawa dan Indonesia), melakukan praproses dokumen hingga melakukan perangkingan n-gram yang didapat dari setiap dokumen. Seluruh profil n-gram dan frekuensinya kemudian disimpan ke dalam basis data di dua tabel yang terpisah, yaitu jawagram untuk profil Bahasa Jawa dan indogram untuk profil Bahasa Indonesia.

Tabel profil n-gram sebuah bahasa menyimpan 3 *fields*, yaitu: ID, karakter, dan frekuensi. Kolom 'ID' ditetapkan sebagai kunci utama (*primary key)*, kolom 'karakter' berisi sekuens n-gram yang telah terbentuk, dan kolom 'frekuensi' menunjukkan nilali CF atau jumlah kemunculan n-gram tersebut. Di tabel profil, sekuens n-gram sudah diurutkan sesuai dengan frekuensinya urut dari nilai tertinggi sampai ke rendah (*descending*), di mana semakin banyak frekuensi dari suatu karakter, maka karakter tersebut akan diletakkan pada ranking yang semakin tinggi. Contoh potongan basis data profil n-gram Bahasa Indonesia ditunjukkan oleh Gambar 4.

Sistem yang dibangun memiliki antar muka yang memudahkan pengguna untuk menambah dokumen sampel profil bahasa, seperti ditunjukkan oleh Gambar 5. Melalui antar muka ini, pengguna dapat memasukkan dokumen dengan format .txt, .doc, .docx, ataupun .pdf untuk profil bahasa yang diinginkan. Dokumen tersebut akan disimpan dalam sebuah folder kemudian dikonversi secara otomatis ke dalam format .txt, dan dilakukan praproses awal termasuk penghilangan seluruh angka dan tanda baca yang ada pada teks dokumen masukan tersebut dan mengganti semua spasi yang ada pada dokumen dengan karakter '_'. Tahap selanjutnya adalah pembuatan profil n-gram dari teks yang telah dipilih user dengan urutan semakin atas adalah karakter yang memiliki frekuensi semakin banyak pula. Masing-masing jenis n-gram akan diambil 100 sekuens dengan frekuensi terbanyak, kemudian ke-400 n-gram tersebut digabungkan ke dalam sebuah senarai dan diurutkan kembali berdasarkan frekuensi terbanyak. Dari hasil pengurutan ini akan diambil 100 sekuens n-gram dengan ranking teratas.

indo_id	indo_char	indo_frek ▼ 1
1	an	6517
2	n_	4260
3	an_	3704
4	a_	3697
5	ng	3049
6	i_	2976
7	_d	2857
8	da	2700
9	er	2665
10	en	2321

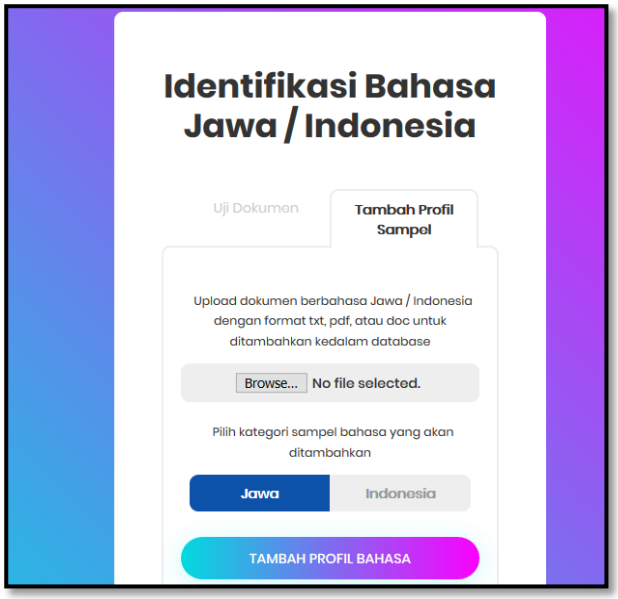
Gambar 4 Contoh tabel di basisdata untuk profil kategori bahasa Indonesia

Dari Gambar 5 tersebut bisa dilihat bahwa antarmuka sistem ini terdiri dari dua menu utama, yakni menu penambahan dokumen untuk pembentukan profil bahasa, dan menu untuk mengidentifikasi bahasa.

Untuk pengujian, sistem ini menyediakan antar muka untuk memasukkan dokumen uji, dan seperti halnya pada pembuatan profil, dokumen uji dapat dimasukkan dengan format .txt, .doc, .docx, ataupun .pdf. Setelah itu, pengguna dapat menekan tombol "Identifikasi Bahasa" untuk mengidentifikasi bahasa dokumen uji. Pada penelitian ini, selain dokumen berbahasa Jawa dan Indonesia, juga diujikan dokumen berbahasa lain seperti bahasa Inggris dan bahasa Melayu. Antarmuka untuk memasukkan dan mengidentifikasi dokumen uji ditunjukkan oleh Gambar 6.

Karena proses identifikasi bahasa memerlukan waktu, maka dirancanglah sebuah *progress bar* seperti ditunjukkan di Gambar 7. *Progress bar* tersebut dilengkapi dengan informasi pada bagian bawah yang menunjukkan tahapan proses yang sedang berjalan. Hal ini diperlukan untuk mencegah pengguna mengira sistem berhenti bekerja karena tidak menampilkan umpan balik apapun, saat sistem masih berusaha menyelesaikan proses identifikasi. Tahapan proses tersebut meliputi:

- Normalisasi dokumen uji
- Pembentukan n-gram dokumen uji,
- Perankingan 100 n-gram teratas dari dokumen uji,
- Penghitungan OOP(D_x, D_{Jv}),
- Penghitungan OOP(D_x, D_{Indo}),
- Menampilkan hasil

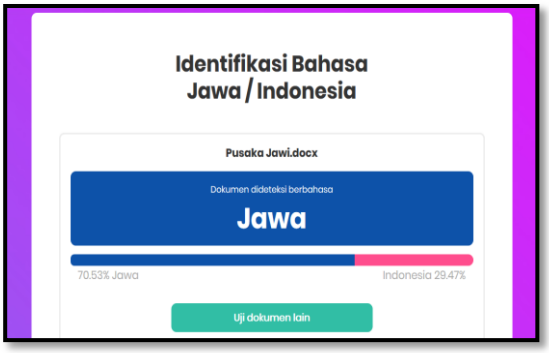


Gambar 5 Antarmuka untuk penambahan data latih

Gambar 6. Antarmuka menu identifikasi bahasa

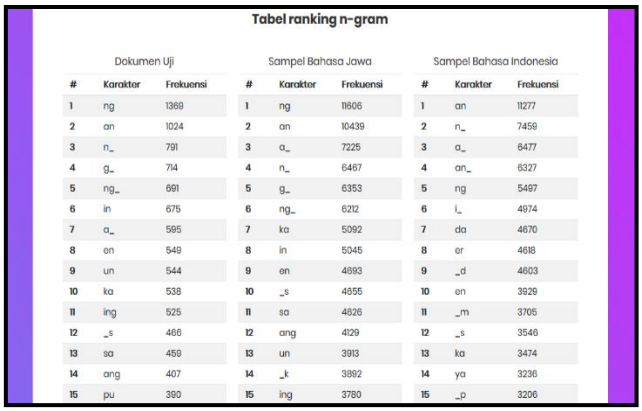
Gambar 7. *Progress bar* identifikasi bahasa dokumen uji

Setelah proses identifikasi selesai, hasilnya akan ditampilkan dalam sebuah jendela dialog yang menunjukkan bahasa yang memiliki jarak terkecil,. Di bawah jendela ini ada batang yang nampak mirip seperti *progress bar* namun mengambarkan hasil perhitungan prosentasi selisih nilai OOP dokumen uji terhadap kedua profil kategori bahasa, dengan bahasa Jawa di sebalah kiri dan bahasa Indonesia di sebelah kanan. Nilai prosentase selisih diproyeksikan ke batang prosentase ini untuk memudahkan pengguna melihat hasil luaran sistem dengan cepat. Gambar 8 menampilkan hasil akhir pengujian dokumen uji yang berjudul Pusaka Jawi. Nama dokumen uji juga ditampilkan di atas hasil kesimpulan identifikasi.



Gambar 8 Antarmuka hasil identifikasi sistem terhadap dokumen uji berjudul Pusaka Jawi.

Selain hasil akhir identifikasi, sistem juga menampilkan rincian 100 n-gram teratas dari dokumen uji, dari profil Bahasa Jawa, dan dari profil Bahasa Indonesia. Tujuan dari ditampilkannya profil ini adalah untuk memperkuat analisis hasil identifikasi bagi peneliti. Jika pengguna ingin mengetahui profil n-gram dari dokumen uji, bisa juga melihat menu ini. Gambar 9 menampilkan contoh dari daftar profil n-gram dari kedua profil kategori bahasa dengan dokumen uji Pusaka Jawi.



Gambar 9. Hasil n-gram dokumen uji, profil Jawa, dan profil Indonesia

B. Pengujian dan Evaluasi Sistem

Dalam penelitian ini, pengujian dilakukan dengan menggunakan 67 dokumen yang berbeda dari dokumen yang dipergunakan sebagai dokumen profil bahasa.

Jumlah dokumen uji berkisar 31 dokumen berbahasa Jawa dan 26 dokumen berbahasa Indonesia, 5 dokumen berbahasa Inggris, dan 5 dokumen berbahasa Melayu. Dokumen yang diujikan memiliki variasi panjang yang sangat mencolok dari dokumen yang terdiri dari 1 kalimat sepanjang 10 kata saja sampai dokumen riil sepanjang 2802 kata. Tujuannya adalah untuk melihat sebagus apakah metode yang diterapkan dapat mengidentifikasi dokumen pendek.

TABEL II
DATA STATISTIK DOKUMEN UJI

Sumber	Jw	Ind	Ing	Mly
wikipedia.org	-	9	1	5
thegorbalsla.com	-	1	-	-
tribunnews.com	-	1	-	-
pengertiandefinisi.com	-	1	-	-
edu.gcfglobal.org	-	5	-	-
msn.com	-	1	-	-
artikelsiana.com	-	1	-	-
tirto.id	-	1	-	-
wordpress.com	-	1	-	-
histori.id	-	4	1	-
sastra.org	31	1	-	-
howstuffworks.com	-	-	1	-
psychologytoday.com	-	-	1	-
jurnal	-	-	1	-
Total	**31**	**26**	**5**	**5**

Hasil identifikasi dokumen-dokumen tersebut dapat dilihat pada Tabel 2. Sesirah kolom sebelah kiri menunjukkan bahasa dokumen uji, sedangkan sesirah kolom bagian atas menunjukkan hasil identifikasinya. Hasil ini dipilah mejadi 2 bagian, dimana bagian kiri menampilkan hasil bagi dokumen pendek dengan panjang kata kurang dari 100, sedangkan bagian kanan menunjukkan hasil identifikasi di dokumen dengan panjang lebih dari 100 kata.

Pada Tabel 3 dapat dilihat bahwa terdapat beberapa kesalahan hasil identifikasi yang ditunjukkan pada pengelompokan di kategori bahasa lainnya. Sebagian besar kesalahan identifikasi terjadi pada dokumen dengan panjang kurang dari 100 kata. Besarnya persentase tersebut disebabkan jumlah kata yang ada pada dokumen terlalu sedikit sehingga menyebabkan profil n-gram yang terbentuk tidak memiliki frekuensi yang cukup untuk dibandingkan dengan dengan profil n-gram kategori bahasa.

TABEL III
MATRIKS HASIL PENGUJIAN YANG DIBEDAKAN DALAM 2 KELOMPOK

Jml Kata	**< 100**				**>=100**				**Σ**
Bahasa	**J**	**I**	**L**	**Σ**	**J**	**I**	**L**	**Σ**	
Jawa (J)	**1**	0	3	**4**	**27**	0	0	**27**	**31**
Indonesia (I)	0	**2**	2	**4**	0	**22**	0	**22**	**26**
Lainnya (L)	0	0	0	**0**	1	4	**5**	**10**	**10**
TOTAL DOKUMEN UJI									**67**

Hasil pengujian lengkap terhadap seluruh dokumen uji dapat dilihat pada Tabel 4. Selisih jarak n-gram terkecil diberikan tulisan tebal untuk menunjukkan bahwa nilai tersebut menyebabkan dokumen diidentifikasikan sebagai dokumen dari bahasa pada sesirah kolomnya. Baris-baris pada kolom "Hasil Identifikasi" yang diberi arsir abu-abu menunjukkan hasil identifikasi yang kurang tepat.

Pada Tabel 4. terdapat kolom "selisih jarak n-gram" yang memiliki sub-kolom "Indonesia" dan "Jawa". Kolom tersebut untuk membandingkan selisih jarak antara n-gram dokumen uji dengan n-gram sampel bahasa Jawa dan n-gram dokumen uji dengan n-gram sampel bahasa Indonesia. Apabila pada kolom Indonesia memiliki angka lebih kecil dibandingkan dengan kolom Jawa, maka hasil identifikasi bahasa dari dokumen uji tersebut adalah bahasa Indonesia, begitu pula sebaliknya. Pada kolom tersebut kita juga bisa membandingkan selisih keduanya dan dapat kita ketahui bahwa semakin kecil jumlah kata pada dokumen uji, maka selisih jarak antara kedua kategori tersebut semakin kecil pula (baris 1-5). Begitu pula sebaliknya semakin banyak kata pada dokumen uji, maka selisih jarak antara kedua kategori tersebut semakin besar. Kondisi ini terjadi oleh karena jumlah variasi n-gram yang terbentuk berbanding lurus dengan jumlah kata pada dokumen, sehingga dokumen dengan jumlah kata yang sedikit, sekuens n-gram yang dihasilkan tidak cukup untuk menentukan kedekatan dengan sebuah profil bahasa.

TABEL IV
HASIL UJI SISTEM DENGAN KASUS-KASUS MENARIK

Dok ID	**Jml Kata**	**Bhs Dok**	**Selisih jarak n-gram**		**Presentase selisih**	**Hasil Identifi-kasi**
			Jawa	**Ind**		
1	10	Ind	9353	**9351**	0,010692	Lainnya
2	10	Jawa	**8771**	8987	1,21635	Lainnya
3	20	Ind	8672	**8514**	0,91935	Lainnya
4	22	Jawa	**7536**	7924	2,50970	Lainnya
5	38	Jawa	**7534**	7919	2,49142	Lainnya
6	50	Ind	7421	**6176**	9,15643	Indonesia
7	53	Jawa	**6213**	7288	7,96237	Jawa
8	99	Ind	6959	**4938**	16,98747	Indonesia
9	103	Jawa	**5483**	6496	8,456465	Jawa
10	152	Jawa	**6949**	7745	5,41717	Jawa
11	199	Ind	6269	**5156**	9,74179	Indonesia
12	210	Jawa	**3964**	6503	24,25718	Jawa
13	230	Ind	6550	**5249**	11,02635	Indonesia
…	…	…	…	**…**	…	…
57	2802	Ind	5272	**2941**	28,38183	Indonesia
58	216	*Mly*	6132	**3706**	24,65948	Indonesia
59	371	*Ing*	**8067**	8362	1,795605	Lainnya
60	542	*Ing*	**7906**	7914	0,05056	Lainnya
61	581	*Mly*	5622	**2815**	33,27011	Indonesia
62	627	*Ing*	**8167**	8203	0,21991	Jawa
63	731	*Mly*	**7314**	7380	0,44916	Lainnya
64	988	*Mly*	6587	**4763**	16,07048	Indonesia
65	1220	*Ing*	**7311**	7982	4,387628	Lainnya
66	1716	*Ing*	**7541**	7925	2,48286	Lainnya
67	1758	*Mly*	5491	**3567**	21,24089	Indonesia

Keterangan:

- Dok ID: Dokumen ID, Jml Kata: Jumlah Kata
- Ind: Indonesia, Ing: Inggris, Mly: Melayu

Dalam kesempatan ini diujikan juga dokumen berbahasa lain, yakni bahasa Inggris dan bahasa Melayu yang sangat dekat dengan bahasa Indonesia. Ada 1 dokumen berbahasa Inggris yang teridentifikasi sebagai bahasa Jawa, namun sisanya diidentifikasi sebagai bahasa 'lainnya' yang pada dasarnya sudah benar. INi disebabkan karena n-gram pada dokumen uji tersebut memiliki kemiripan dengan n-gram dalam bahasa Jawa. Beberapa contoh n-gram karakter yang menempati ranking atas pada dokumen berbahasa Inggris yang memiliki kemiripan dengan karakter n-gram berbahasa Jawa antara lain: "_in", "ing", dan "an".

Berbeda dari dokumen berbahaa Inggris, percobaan dokumen berbahasa Melayu menunjukkan angka persentasi pengenalan yang tinggi sebagai bahasa Indonesia, yang sebenarnya kurang tepat. Ini disebabkan dokumen berbahasa Melayu memiliki n-gram yang jauh lebih mirip dengan n-gram pada database bahasa Indonesia, karena bahsa Melayu merupakan induk bahasa Indonesia. Beberapa contoh karakter n-gram dokumen yang memiliki ranking atas pada dokumen berbahasa Melayu yang memiliki kemiripan dengan karakter n-gram berbahasa Indonesia antara lain: "an", "ng", dan "ah".

Selain jumlah kata unik pada dokumen uji, jumlah data sampel bahasa Indonesia dan bahasa Jawa serta variasinya juga sangat mempengaruhi luaran sistem identifikasi ini, karena profil sampel bahasa Indonesia dan Jawa merupakan pembanding untuk menentukan selisih jarak dari n-gram dokumen uji. Semakin banyak dan semakin bervariasi profil sampel, maka dapat menambah akurasi sistem.

Berdasarkan pengujian terhadap 67 dokumen yang hasilnya telah ditampilkan di tabel III dan IV, maka kinerja sistem pun perlu dievaluasi dengan metriks akurasi dan presisi. Akurasi merupakan metrik yang menghitung jumlah total dokumen yang teridentifikasi dengan benar dibagi dengan jumlah total dokumen. Jika menggunakan prespektif klasifikasi maka penghitungan akurasi didasarkan pada matrik kebingungan (confusion matrix) dimana jumlah benar positif ditambahkan dengan jumlah negative benar kemudian dibagi dengan jumlah total dokumen yang diujikan. Penghitungan akurasi dilakukan berdasarkan persamaan 3.

$$Akurasi(D) = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Berdasarkan rumus di atas dan tabel III, maka nilai akurasi yang diperoleh sistem bisa dihitung, yakni 57/67 = 0.85. Nilai akurasi ini tidaklah tinggi yang disebabkan oleh dua faktor. Yang pertama, jumlah dokumen uji antar kelasnya cukup seimbang. Faktor lainnya adalah diujikannya dokumen dengan kategori bahasa yang tidak termasuk dalam kategori yang ditentukan, namun karena kedekatan bahasa Melayu kepada bahasa Indonesia, maka prosentasi identifikasi masuk ke bahasa Indonesia sangatlah tinggi. Dan tentu saja ini kami perhitungkan sebagai kesalahan.

Dalam perspektif klasifikasi, Akurasi dipandang sebagai metrik yang kurang tepat karena kecenderungan biasnya dan memberi nilai akurasi tinggi pada data yang tidak seimbang saat jumlah *false negative*-nya tinggi. Sebagai alternatifnya, maka ukuran Presisi digunakan. Presisi akan menghitung jumlah prediksi benar (TP) dibagi dengan jumlah total prediksi (TP + FP). Persamaan untuk Presisi bisa dilihat di Persamaan 4.

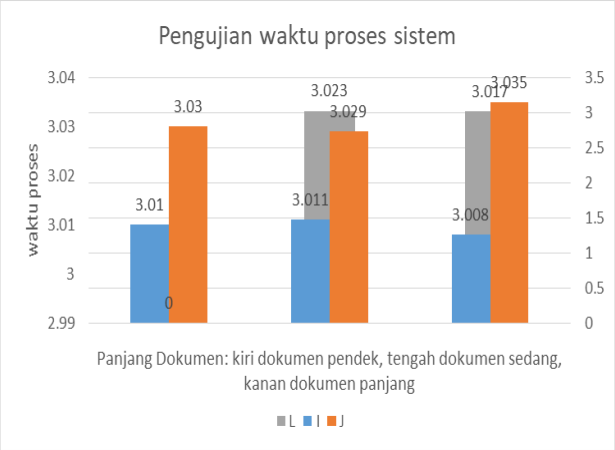
$$Presisi(D) = \frac{TP}{TP+FP} \quad (4)$$

Dalam tahap evaluasi ini, penghitungan presisi diberlakukan per kategori bahasa, berhubung identifikasi diberlakukan pada 3 kelas bahasa yang berbeda. Berdasarkan tabel III, maka nilai Presisi untuk identifikasi dokumen Jawa, Indonesia dan 'lain'-nya bisa dihitung sebagai berikut:

- Presisi(D_{JV}) = 28/28+1 = 0.96
- Presisi(D_{Indo}) = 24/24+4 = 0.86
- Presisi(D_{L}) = 5/5+5 = 0.5

Dari perhitungan nilali Presisi, terlihat dengan jelas, bahwa sistem mengidentifikasi dokumen berbahasa Jawa jauh lebih tepat dibandingkan dengan dokumen berbahasa Indonesia. Kesalahan identifikasi bahasa Jawa terjadi pada dokumen yanga sangat pendek dengan jumlah kata kurang dari 50. Sedangkan nilai Presisi kategori bahasa Indonesia jauh di bawah nilai Presisi bahasa Jawa, dikarenakan diujikannya dokumen berbahasa Melayu (Malaysia) yang teridentifikasi dalam kategori bahasa Indonesia. Keterkaitan ini jelas dan logis karena tidak ada satupun dokumen uji berbahasa Melayu yang diidentifikasi sebagai bahasa Jawa. Nilai presisi kategori bahasa lainnya tegolong cukup rendah yang hanya mencapai 0.50 karena dari 10 dokumen yang diujikan, 5 diantaranya (50%) adalah dokumen berbahasa Melayu yang 40% disalah-kategorikan sebagai bahasa Indonesia. Sumbagan kesalahan identifikasi terjadi pada sebuah dokumen berbahasa Inggris yang teridentifikasi sebagai dokumen berbahasa Jawa.

Pengujian terhadap waktu proses juga dilakukan untuk mengetahui rata-rata waktu proses. Dengan menggunakan dokumen uji yang sama, maka dilakukan pengamatan terhadap waktu proses identifikasi antara dokumen pendek dengan jumlah kata dibawah 100, dokumen sedang, dengan panjang antara 100-1.000 kata, dan dokumen panjang dengan jumlah kata lebih dari 1000. Hasil pengujian pemrosesan waktu ditampilkan di gambar 10 berikut ini.



Gambar 10. Pengujian waktu proses identifikasi dimana L merujuk pada rata-rata waktu identifikasi bahasa lainnya, I pada dokumen berbahasa Indonesia, dan J untuk bahasa Jawa.

Dari gambar 10, bisa dilihat bahwa waktu proses identifikasi bahasa relatif konstan yakni 3 detik. Panjang-pendek dokumen tidak terlalu mempengaruhi waktu proses, dikarenakan jumlah profil bahasa yang dibandingkan adalah stabil, yakni 100 baik bagi dokumen panjang maupun pendek. Rata-rata waktu terpendek adalah untuk deteksi dokumen berbahasa Indonesia dengan panjang lebih dari 1000 kata yakni 3.008 detik, sedangkan rata-rata waktu tertinggi terjadi saat identifikasi kategori bahasa lainnya dengan lama 3.035 detik. Selisih waktu terendah dan tertinggi hanya mencapai 0.027 detik. Selain itu, yang dibandingkan adalah profil n-gram bukan token atau kata sehingga panjang dokumen tidak memiliki korelasi langsung dengan waktu proses, kecuali jika jumlah profil dokumen dan bahasa ditentukan secara proporsional dengan panjang dokumen.

VI. KESIMPULAN

Dari hasil penelitian yang dilakukan, maka dapat dikatakan bahwa karakter n-gram dalam proses identifikasi bahasa memiliki nilai total akurasi 85,07463%. Sedangkan nilai Presisi kategori bahasa Jawa mencapai 0.96, dan 0.86 bagi kategori bahasa Indonesia. Hal ini membuktikan bahwa n-gram karakter dengan profil 100 n-gram tertinggi dapat diterapkan dalam proses pengidentifikasian bahasa. Tingkat presisi identifikasi meningkat saat dokumen memiliki panjang lebih dari 50 kata. Namun untuk mendeteksi sebuah kalimat atau dokumen pendek yang terdiri dari 10-40 kata, sistem ini masih mengalami kesulitan. Untuk meningkatkan presisi pengenalan pada dokumen pendek atau pada sebuah kalimat panjang, diperlukan untuk mengabungkan n-gram dengan profil lainnya, seperti stop-word, misalnya. Pengujian terhadap waktu proses menunjukkan bahwa waktu proses identifikasi bahasa relatif stabil dan membutuhkan 3 detik. Panjang-pendek dokumen tidak mempengaruhi waktu proses dikarenakan yang dibandingkan adalah profil n-gram dengan jumlah yang tetap.

REFERENSI

[1] M. Majlis, "Yet Another Language Identifier," dalam *Proceedings of the EACL 2012 Student Research Workshop*, Avignon, France, 2012.

[2] J. Gonzalez-Dominguez, I. Ignacio Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez dan P. J. Moreno, "Automatic Language Identification Using Long Short-Term Memory Recurrent Neural Networks," dalam *INTERSPEECH*, 2014.

[3] M. Lui, J. H. Lau dan T. Baldwin, "Automatic Detection and Language Identification of Multilingual Documents," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 27-40, 2014.

[4] W. B. Cavnar dan J. M. Trenkle, "N-Gram-Based Text Categorization," dalam *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, USA, 1994.

[5] T. Baldwin dan M. Lui, "Language Identification: The Long and the Short of the Matter," dalam *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, California, 2010.

[6] l. Ferrer, N. Scheffer dan E. Shriberg, "AComparisonof Approaches for Modeling Prosodic Features in Speaker Recognition," dalam *International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[7] D. Martinez, E. Lleida, A. Ortega dan A. Miguel, "Prosodic features and formant modeling for an ivectorbased language recognition system," dalam *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[8] N. Dehak, P. A. T. Carrasquillo, D. Reynolds dan R. Dehak, "Language Recognition via Ivectors and Dimensionality Reduction," dalam *INTERSPEECH*, Florence, Italy, 2011.

[9] H. Takcı dan I. Soğukpınar, "Letter Based Text Scoring Method for Language Identification," dalam *ADVIS: International Conference on Advances in Information Systems*, Izmir, Turkey, 2004.

[10] H. Takçı dan E. Ekinci, "Minimal Feature Set in Language Identification and Finding Suitable Classification Method with It," *Procedia Technology*, vol. I, pp. 444-448, 2012.

[11] A. Selamat dan N. Akosu, "Word-Length Algorithm for Language Identification of Under-Resourced Languages," *Journal of King Saud University – Computer and Information Science*, vol. 28, pp. 457-469, 2014.

[12] N. Sarma, S. R. Singh dan D. Goswami, "Word Level Language Identification in Assamese-Bengali-Hindi-English Code-Mixed Social Media Text," dalam *International Conference on Asian Language Processing (IALP)*, Bandung, Indonesia, 2018.

[13] A. Selamat, "Improved N-grams Approach for Web Page Language Identification," dalam *Transactions on Computational Collective Intelligence V*, N. Nguyen, Penyunt., Heidelberg, Springer, 2011, pp. 1-26.

[14] L. D. Krisnawati dan A. W. Mahstama, "A Javanese Syllabifier based on Its Orthographic System," dalam *2018 International Conference on Asian Language Processing (IALP)*, Bandung, Indonesia, 2018.

[15] J. Garg, V. Grupta dan M. Jindal, "A Survey of Language Identification Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence*, vol. 6, no. 4, pp. 388-399, 2014.