ISSN 2338-1523
E-ISSN 2541-576X

Volume 10 No. 3
Desember 2022

NORMALISASI KATA BAHASA BATAK MANDAILING MENGGUNAKAN METODE *LEVENSHTEIN DISTANCE*

Sindi Maulia[1)*], Erni Rouza[2)], Luth Fimawahib[3)], Imam Rangga Bakti[4)], Satria Riki Mustafa[5)]

[1,2,3,4,5)] Teknik Informatika, Universitas Pasir Pengaraian, Rokan Hulu, Riau
email: sindimaulia83@gmail.com[1)], ernirouzait@gmail.com[2)], luthfimawahib@gmail.com[3)], imamranggabakti@gmail.com[4)], satriarikimustafa@gmail.com[5)]

Abstrak

Batak Mandailing merupakan suatu masyararkat Indonesia yang menggunakan bahasa Batak Mandailing daerah yang mereka tempati terletak pada kabupaten tapanuli setalatan (Sayur Matinggi) serta Kabupaten Mandailing Natal yang meliputi daerah Panyabungan, Siabu, Hutan Pungkut, Panyabungan dan lainnya, di Provinsi Sumatera Utara. Pada penelitian ini dalam normalisasi kata bahasa batak mandailing digunakan metode *Levenshtein Distance*, metode ini adalah matriks perbandingan dalam mengukur perbedaan diantara dua urutan. *Levenshtein Distance* sering dugunakan pada perbandingan diantara dua urutan *String* yang membantu masalah dalam memperbaiki setiap kesalahan pada ejaan didalam kata, pada pembuatan aplikasi untuk normalisasi kata Bahasa batak mandainling ini digunkan Bahasa pemograman PHP dan MySQL sebagai basis datanya. Hasil dari penggunaan metode *Levenshtein Distance* ini semakin banyak kata dasar yang dimiliki maka hasilnya akan semakin bagus dan juga jika melebihi batasan String nya maka kata tidak akan di temukan dikarenakan batasan masalah pada penelitian ini yaitu jarak *Levenshtein Distance* yang digunakan dalam pengubahan yang memiliki nilai 0 sampai kurang dari 3 *string* saja. Dari hasil pengujian diperoleh akurasi sebesar 84% sehingga dapat disimpulkan bahwa aplikasi ini dapat diterima dengan baik.

Kata Kunci : Bahasa Batak Mandailing, Levenshtein Distance, MySQL, PHP.

Abstract

The Mandailing Batak are an Indonesian people who speak the Mandailing Batak language, the area where they live is located in the Tapanuli Selatan District (Sayur Matinggi) and the Mandailing Natal Regency which includes the Panyabungan, Siabu, Pungkut Forest, Panyabungan and others, in North Sumatra Province. In this study, in normalizing the words of the Batak Mandailing language using the Levenshtein Distance method, this method is a comparison matrix in measuring the difference between two orders. Levenshtein Distance is often used in comparisons between two string sequences which help with problems in correcting any spelling errors in words, in making applications for normalizing words in the Batak Mandainling language the PHP and MySQL programming languages are used as the database. The result of using the Levenshtein Distance method is that the more basic words you have, the better the results will be and also if it exceeds the String limit, the word will not be found because of the problem in this study, namely the Levenshtein Distance used in the conversion which has a value of 0 to less of 3 strings only. From the test results obtained an accuracy of 84% so it can be concluded that this application can be well received.

ISSN 2338-1523
E-ISSN 2541-576X

Volume 10 No. 3
Desember 2022

Keywords: *Batak Mandailing ,Levenshtein Distance, MsQL, PHP.*

PENDAHULUAN

Indonesia adalah sebuah negara yang sangat kaya akan ragam dan budaya. Pada setiap daerah Indonesia memiliki ragam dan budayanya masing masing. Seperti rumah adat, tarian tradisonal, pakaian adat, alat musik dan Bahasa khas daerah. Dari banyaknya Bahasa di Indonesia ada Bahasa bahasa Batak Mandailing. Bahasa Batak Mandailing merupakan lingua franca (bahasa pergaulan/bahasa sehari-hari) untuk masyarakat Tapanuli Selatan (TAPSEL) Kota Padang Sidimpuan. Bahasa Batak Mandailing dipergunakan sebagai Bahasa komunikasi pada pergaulan baik pada sahabat, keluarga ataupun untuk keperluan dan kepentingan lainnya yang tidak formal [1].

Metode *Levenshtein Distance* adalah suatu algoritma yang bisa melakukan pengukuran kesamaan antara string. Penggabungan antara algoritma *Levenshtein Distance* dan teknik *Scraping* ini yang akan membantu masyarakat dalam menulis Bahasa Batak Mandailing sehingga tidak mengalami kesalahan makna dan penulisan [2].

Berdasarkan permasalahan yang telah dikemukakan, peneliti mencoba menerapkan metode *Levenshtein Distance* untuk Bahasa Batak Mandailing sehingga diharapkan dapat membantu pihak sekolah dalam proses pembelajaran bahasa Batak. Oleh karenanya bisa dilakukan penelitian dengan judul "Normalisasi Kata Bahasa Batak Mandailing Menggunakan Metode *Levenshtein Distance*".

NLP (*Natural Language Processing*) merupakan sebuah proses dalam pembuatan model komputasi dari bahasa sehingga dapat terjadi interaksi diantara komputer dan manusia dengan perantara bahasa alami yang dipakai manusia. NLP (*Natural Language Processing*) memodelkan sebuah pengetahuan pada bahasa, baik dari segi kata, agar kata-kata bergabung menjadi sebuah kalimat dan konteks kata dalam kalimat [5].

Text mining merupakan penambangan data yang berbentuk teks dimana biasanya sumber data di dapat dari dokumen, serta tujuannya untuk mencari kata yang bisa mewakili dari isi dokumen sehingga bisa dilakukannya analisa yang terhubung antar dokumen. *Text mining* juga suatu penerapan teknik dan konsep data mining untuk melakukan pencarian pola pada teks, yaitu suatu proses analisis teks guna mencari sebuah informasi yang bermanfaat untuk suatu tujuan tertentu [8].

Tahapan Text Mining klasifikasi merupakan *pattern discovery* (tahap penemuan pola) tetapi secara lengkap proses pada text mining dibagi menjadi 3 tahap, yaitu [10] :

a. *Text Prepocessing*
b. *Text Transformation* (*Feature Generation*)
c. *Pattern Discovery*

Normalisasi merupakan suatu file terdiri dari beberapa elemen group yang berulang perlu diorganisasikan kembali. Proses dalam perorganisasian file dengan melakukan penghilanagn terhadap elemen group yang berulang atau langkah proses dalam penyederhanaan suatu hubungan antar elemen data dalam *record* (*tuple*) ini yang disebut juga dengan normalisasi. Normalisasi sangat banyak dilakukan untuk perubahan dalam bentuk database dari sebuah struktur pohon ataupun struktur jaringan menjadi sebuah struktur yang berhubungan. [9].

ISSN 2338-1523
E-ISSN 2541-576X

Volume 10 No. 3
Desember 2022

Levenshtein Distance merupakan matriks perbandingan yang bisa mengukur perbedaan anatara dua urutan. *Levenshtein Distance* juga sering digunkan didalam membandingankan diantara dua urutan String yang berguna untuk masalah memperbaiki kesalahan eja dalam kata [5]. *Levenshtein Distance*, atau sering disebut juga sebagai *edit distance*, merupakan matriks atau pengukuran yang dihasilkan oleh perhitungan jumlah perbedaan yang dimiliki oleh dua string. *Levenshtein Distance* yang dimiliki oleh dua string adalah jumlah minimum perubahan yang diperuntukkan sebagai mengganti sebuah string terhadap string lain, dengan operasi hapus (*delete*), tambah (*insert*), atau pergantian karakter (*substitute*) terhadap suatu karakter [6].

Batak Mandailing merupakan masyarakat yang penggunaan bahasanya adalah bahasa Batak Mandailing serta daerah tempat tinggal dari suku Batak Mandailing ini terletak pada Kabupaten Tapanuli Selatan (Sayur Matinggi) dan Kabupaten Mandailing Natal (Siabu, Panyabungan, Kotanopan, Huta Pungkut, dll), di Sumatera Utara [8].

MySQL merupakan *software* sistem manajemen yang dibasiskan oleh basis data SQL (bahasa Inggris: *database management system*) atau DBMS yang multi-user, multithread, dengan sekitar 6 juta instalasi pada keseluruhan dunia. MySQL AB membuat MySQL tersedia utnuk perangkat lunak gratis di naungi lisensi GPL (*General Public License*), tapi merekan juga melakukan penjualan di bawah lisensi komersial terhadap kasus kasu yang mana penggunaanya tidak cocok terhadap GPL [19].

PHP secara umum adalah Bahasa pemograman script yang mana membuat dekumen dengan HTML secara *on the fly* yang mana nantinya akan dieksekusi pada *server web*, dokumen HTML yang dihasilkan pada suatu aplikasi bukanlah HTML yang dibuat dengan menggunakan *text editor* atau editor HTML. Yang mana dikenal juga sebagai bahasa pemrograman *server side* [11].

HASIL DAN PEMBAHASAN

Tahapan awa pada Algoritma *Levenshtein Distance* yaitu melajyjan perhitungan jarak string diantara sumber dan target, untuk contoh disini kata “AU” sebagai target dan “AU” sebagai sumbernya. Perhitungan dilakukan dengan inisisasi urutan karakter terhadap setiap string seperti yang ditabelkan pada tabel dibawah berikut :

Tabel 1. Jarak D (1,1) Kata AU

		A	U
	0	1	2
A	1	0	
U	2		

Seetelahnya dilakukan perhitungan jarak terhadap karakter ke-1 string sumber bersama karakter ke-2 terhadap string target serta diketahui bahwa dibutuhkannya operasi penyisipan terhadap karakter “a” oleh string sumber, sehingga nilai yang diberikan bisa sesuai dengan rumus operasi penyisipan terhadap rumus 2, yaitu $D(1, 2) = D(1, 2 - 1), +1$. sehingga nilai yang diberikan pada D(1, 2) = D(1, 1) + 1 yang bernilai D (1, 2) = 0+1 = 1. Kemudian nilai matrik pada D (1, 2) diisi dengan nilai 1 seperti yang ditabelkan pada tabel dibawah berikut :

Tabel 2. Jarak D (1,1) Kata AU

		A	U
	0	1	2

ISSN 2338-1523
E-ISSN 2541-576X

Volume 10 No. 3
Desember 2022

A	1	0	1
U	2		

Untuk hitungan jarak terhadap karakter ke-1 string sumber dengan string target ke-2 karena huruf target dan sumber sama maka jarak antara karakter ke-1 string ke-1 dan string target ke-2 adalah 0, maka untuk nilai matrik pada D(2,2) diisi dengan nilai 0 seperti pada tabel 3

Tabel 3. Hasil Perhitungan Matrik

		A	U
	0	1	2
A	1	0	1
U	2	1	0

SIMPULAN

Dari hasil pengujian, keakuratan proses normalisasi ditentukan pada kelengkapan kata dasar semakin banyak kata dasar maka akan besar akurasi dalam normalisasi kata bahasa Batak Mandailing, Berdasarkan hasil perhitungan tikat akurasi dari 230 kata uji menunjukkan persentase yaitu 21,73% dengan pengujian sebanyak 50 kali dari 230 kosa kata.

UCAPAN TERIMA KASIH

Terimaksih kepada rekan rekan yang turut membantu serta mendukung dalam penyelesaiannya jurnal ini.

DAFTAR PUSTAKA

[1] A. K. Arsyad, B. Pramono, Isnawaty,M. Yamin, and Ihsan, "Implementasi *Levenshtein Distance* Pada Aplikasi Pencarian Barang Di Berbagai E-Marketplace Menggunakan Teknik Web Scraping," *Semin. Nas. APTIKOM 2019*, vol. 1, no. 1, pp.512–519, 2019.

[2] A. N. Rohman, E. Utami, and S. Raharjo, "Deteksi Kondisi Emosi pada Media Sosial Menggunakan Pendekatan Leksikon dan Natural Language Processing," *Eksplora Inform.*, vol. 9, no. 1, pp. 70–76, 2019,doi:10.30864/eksplora.v9i1. 277.

[3] S. J. Sitompul, "Interferensi Bahasa Batak Mandailing pada Penggunaan Bahasa Indonesia dalam Interaksi Kelas di Kelas VII Madrasyah Tsanawiyah Swasta," *Edukasi Kult. J.Bahasa, Sastra dan Budaya*, vol. 1, no. 2, pp. 99–114, 2015.

[5] E. A. Lisangan, "Natural Language Processing Dalam Memperoleh Informasi Akademik Natural Language Processing Dalam Memperoleh Informasi Akademik Mahasiswa," *J. Temat.*, vol. 1, no. March 2013, pp. 1–9, 2015.

[6] N. H. Ariyani, Sutardi, and R. Ramadhan, "Aplikasi Pendeteksi Kemiripan Isi Teks Dokumen Menggunakan Metode *Levenshtein Distance*," *semanTIK*, vol. Vol 2, no.1, pp. 279–286, 2016.

[10] Wulandari and S. Aprilia, "Sistem Informasi Penjualan Produk Berbasis Web Pada Chanel Distro Pringsewu," *Technol. Accept. Model*, vol. 4, pp. 1–7, 2015.

[11] H. Hasugian and A. N. Shidiq, "Rancang Bangun Sistem Informasi Industri Kreatif Bidang Penyewaan Sarana Olahraga," *Semin. Nas. Teknol. Inf. dan Komun. Terap. 2012*,vol. 2012, no. Semantik 2012, pp. 606–612, 2012.