Terbit *online* pada laman web jurnal : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Pengembangan Korpus Bahasa Minang pada *Spell Error Corpus for Minang Language* (SPEML)

Dewi Soyusiawaty[a,*], *Abdul Fadlil*[b], *Sunardi*[c]

[a]*Program Studi Informatika, Universitas Ahmad Dahlan, Ringroad Selatan, Yogyakarta 55191, Indonesia*
[b,c]*Program Studi Teknik Elektro, Universitas Ahmad Dahlan, Ringroad Selatan, Yogyakarta 55191, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 21 Desember 2024
Revisi Akhir: 23 April 2025
Diterbitkan *Online*: 30 April 2025

KATA KUNCI

Bahasa Minang,
Kesalahan Ejaan,
Korpus,
SPECIL,
SPEML

KORESPONDENSI

E-mail: dewi.soyusiawaty@tif.uad.ac.id

A B S T R A C T

Bahasa Minang merupakan bahasa daerah kelima dengan jumlah penutur terbanyak di Indonesia, namun minim sumber daya linguistik dan teknologi pemrosesan bahasa alami yang mendukung. Keterbatasan ini menyulitkan pengembangan aplikasi seperti mesin penerjemah dan pemeriksa ejaan otomatis. Saat ini hanya tersedia korpus kesalahan ejaan dalam Bahasa Indonesia dengan kesalahan hanya satu karakter pada setiap token. Korpus belum mencakup kesalahan penulisan kata serapan. Selain itu belum ada korpus khusus yang dikembangkan untuk kesalahan ejaan dalam bahasa daerah di Indonesia, termasuk Bahasa Minang. Penelitian ini bertujuan mengembangkan korpus kesalahan ejaan Bahasa Minang, yang dinamakan *Spell Error Corpus for Minang Language* (SPEML). SPEML mencakup kesalahan ejaan sampai dengan tiga karakter dan kesalahan penulisan kata serapan. Pengembangan SPEML melibatkan proses pengumpulan data korpus Bahasa Minang, data kata serapan yang sering digunakan, serta pembentukan korpus kesalahan ejaan. Kesalahan ejaan dibentuk dengan mengacak token secara sistematis pada satu karakter, dua karakter, hingga tiga karakter, disesuaikan dengan panjang token. Hasil penelitian ini berupa SPEML yang mampu mengklasifikasikan tujuh jenis kesalahan ejaan, yaitu: penyisipan karakter, penghapusan karakter, pindah posisi karakter, penggantian karakter, kesalahan tanda baca, kesalahan kata nyata, dan kesalahan penulisan kata serapan. Pengembangan SPEML menjadi langkah awal dalam mendukung pengembangan teknologi pemrosesan bahasa alami untuk bahasa daerah, khususnya Bahasa Minang.

1. PENDAHULUAN

Indonesia sebagai negara kepulauan terbesar di dunia memiliki keanekaragaman hayati, salah satunya adalah kebinekaan bahasa. Hal ini menjadikan Indonesia sebagai negara dengan keberagaman bahasa terbanyak kedua di dunia [1]. Bahasa daerah (tidak termasuk dialek dan subdialek) di Indonesia yang telah diidentifikasi dan divalidasi sebanyak 718 bahasa dari 2.560 daerah pengamatan [2]. Sebagian besar bahasa dikategorikan sebagai bahasa yang terancam punah dan beberapa bahkan telah punah. Kurangnya kesadaran terhadap bahasa daerah sebagai kekayaan intelektual menjadi salah satu alasan yang menyebabkan bahasa daerah ditinggalkan [3].

Bahasa Minang adalah bahasa daerah kelima yang memiliki jumlah penutur terbanyak di Indonesia setelah Jawa, Sunda, Melayu, dan Madura, namun masih minim sumber daya linguistik dan teknologi pemrosesan bahasa alami yang mendukung [4]. Bahasa Minangkabau sebagai bahasa daerah berfungsi sebagai (a) lambang kebangsaan daerah Sumatera Barat dan pendukung perkembangan kebudayaan Minangkabau, (b) lambang identitas daerah Sumatera Barat dan masyarakat Minangkabau sebagai salah satu suku bangsa di Indonesia, dan (c) alat perhubungan dalam keluarga dan masyarakat Minangkabau dalam komunikasi lisan juga komunikasi lisan antaretnis di Sumatera Barat.

Bahasa Minangkabau sebagai salah satu cabang bahasa Melayu Polinesia mempunyai kemiripan yang sangat dekat dengan Bahasa Indonesia, baik kosa-kata, morfem, maupun sintaksis. Penemuan tentang linguistik Bahasa Minangkabau dapat dijadikan penunjang linguistik nusantara pada umumnya dan linguistik Bahasa Indonesia pada khususnya [5]. Jika Iuas

https://doi.org/10.25077/TEKNOSI.v11i1.2025.17-26

penyebaran Bahasa Minangkabau dijadikan patokan maka wilayah penggunaan Bahasa Minangkabau tidak kalah dengan Bahasa Indonesia [6]. Penutur bahasa Minangkabau berada di seluruh pelosok tanah air, bersamaan dengan tempat tinggal para perantau suku bangsa Minangkabau. Tingkat migrasi suku bangsa Minangkabau merupakan yang tertinggi dari seluruh suku bangsa di Indonesia, bahkan ada pameo yang mengatakan bahwa "di setiap keramaian ada warung nasi padang" dan "di sekitar warung nasi padang, banyak orang Padang" [7].

Korpus adalah kumpulan teks atau data bahasa lisan yang terstruktur, dipilih dan dikumpulkan untuk keperluan analisis linguistik. Sumbernya bisa beragam, seperti buku, surat kabar, transkrip percakapan, hingga konten daring. Dalam linguistik komputasional dan pemrosesan bahasa alami, korpus dimanfaatkan untuk mengembangkan dan menguji algoritma, menganalisis pola dan struktur bahasa, serta mengeksplorasi penggunaan bahasa dalam berbagai konteks [8]. Dalam konteks bahasa Indonesia, korpus dapat digunakan untuk mempelajari tata bahasa, sintaksis, wacana, atau untuk mengembangkan teknologi seperti sistem penerjemahan mesin, *text-to-speech*, dan teknologi bahasa lainnya [9].

Korpus dapat dibuat secara manual maupun otomatis, sering kali dilengkapi dengan anotasi linguistik seperti tag *part-of-speech*, label semantik, atau struktur sintaksis untuk mendukung analisis yang lebih mendalam. Beberapa korpus yang dapat diakses publik, seperti korpus baru Bahasa Indonesia dan korpus Wikipedia Bahasa Indonesia, telah digunakan dalam penelitian tentang kesalahan tata bahasa. Namun, hingga kini belum ada korpus resmi atau terbuka yang secara khusus mengidentifikasi kesalahan ejaan dalam bahasa daerah.

Pengembangan korpus khusus untuk mendeteksi kesalahan ejaan merupakan langkah strategis dalam meningkatkan kemampuan alat pemeriksa kesalahan seperti pemeriksa tata bahasa dan perangkat lunak *proofreading*. Dengan korpus khusus, akurasi dan relevansi deteksi kesalahan dapat ditingkatkan, memungkinkan pengembang menciptakan alat yang lebih baik [10]. Kalimat sebagai unit dasar bahasa, terdiri dari kata-kata yang disusun dalam urutan tertentu untuk menyampaikan gagasan atau pemikiran secara utuh. Kalimat dapat berupa struktur sederhana maupun kompleks, terdiri atas satu atau lebih klausa. Perannya sangat penting dalam komunikasi efektif karena memungkinkan penutur dan penulis menyampaikan ide dengan jelas dan teratur. Kajian terhadap sintaksis dan semantik kalimat menjadi kunci dalam memahami struktur dan maknanya [11].

Penelitian ini bertujuan mengembangkan *Spell Error Corpus for Minang Language* (SPEML) yang merupakan korpus kesalahan ejaan dalam Bahasa Minang dengan tujuh jenis kesalahan yaitu kesalahan penambahan karakter, penghapusan karakter, pindah posisi karakter, penggantian karakter, kesalahan tanda baca, kesalahan kata nyata, dan kesalahan penulisan kata serapan dengan pengacakan sampai dengan tiga karakter.

1.1. Korpus

Penelitian mencatat peningkatan tren jumlah studi terkait korpus bahasa di Indonesia. Belum terlalu banyak upaya membangun dan memperluas *dataset* paralel yang kurang terwakili untuk bahasa lokal Indonesia. Sebagian besar artikel masih fokus pada penggunaan korpus, bukan pada pengembangan korpus. Terdapat kebutuhan mendesak untuk pengembangan korpus multibahasa yang mencakup bahasa daerah dan Bahasa Indonesia. Hanya sejumlah kecil kumpulan data berlabel yang tersedia untuk bahasa lokal di Indonesia [12].

Dataset *Named Entity Recognition* (NER) mencakup Bahasa Aceh, Jawa, Minangkabau, dan Sunda [13]. Kumpulan data multibahasa untuk deteksi bahasa kasar dan ujaran kebencian yang melibatkan bahasa Jawa, Sunda, Madura, Minangkabau, dan Musi [14]. Beberapa kumpulan data tersedia untuk bahasa individual, misalnya, analisis sentimen dan terjemahan mesin dalam Bahasa Minangkabau [15] dan klasifikasi emosi dalam Bahasa Sunda [16]. Kajian pengembangan *dataset* MadureseSet merupakan versi digital dari dokumen fisik Kamus Lengkap Bahasa Madura-Indonesia [17]. Korpus paralel berbasis manusia berkualitas tinggi pertama dalam sepuluh bahasa dari Indonesia yang disebut sebagai NusaX terdiri atas Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, dan Toba Batak serta data paralel dalam Bahasa Indonesia dan Bahasa Inggris yang mencakup tugas analisis sentimen dan terjemahan mesin [1].

Bhinneka Korpus yang dihasilkan merupakan korpus paralel pertama untuk lima bahasa lokal Indonesia (Ambonese Malay, Kupang Malay, Beaye, Makassarese, dan Uab Meto). Korpus ini menyediakan pasangan kalimat dalam Bahasa Indonesia dan bahasa lokal. Berbagai jenis korpus digunakan untuk studi linguistik, pengajaran bahasa, studi budaya, dan analisis wacana [18].

Penelitian lainnya yaitu korpus kesalahan ejaan pertama Bahasa Indonesia bernama *Spell Error Corpus for Indonesian Language* (SPECIL) memberikan kontribusi signifikan dengan menciptakan sumber daya yang komprehensif. SPECIL terdiri dari lebih dari 180.000 token dalam 21.500 kalimat, mencakup berbagai jenis kesalahan ejaan seperti kesalahan non-kata, kesalahan kata nyata, dan kesalahan tanda baca. Kesalahan non-kata meliputi kesalahan pertukaran tempat (*transposition error*), kesalahan penggantian kata (*substitution error*), kesalahan penyisipan kata (*insertion error*), kesalahan penghapusan (*deletion error*), kesalahan kata nyata (*real-word*), dan kesalahan tanda baca (*punctuation errors*) [10].

SPECIL dapat digunakan untuk melatih dan menguji berbagai model *Natural Language Processing* (NLP) termasuk pemeriksa ejaan dan model bahasa, untuk meningkatkan akurasi dan efektivitas dalam mengidentifikasi serta memperbaiki kesalahan dalam teks Bahasa Indonesia. Lingkup SPECIL terbatas hanya satu karakter kesalahan pada tiap kata [19]. Penelitian [20] menggunakan dataset SPECIL untuk melakukan perbaikan kesalahan ejaan menggunakan algoritma *Levenstein Distance* dan *Jaro-Winkler*, yang berfokus pada kesalahan ejaan dengan satu karakter. Penelitian ini berhasil menemukan bahwa jumlah N/minimal operasi perubahan dapat memberikan pengaruh terhadap akurasi. Hasil akhir menunjukkan bahwa *Levenstein Distance* dan *Jaro Winkler*, memberikan akurasi tertinggi yang sama yaitu 99,52% dengan N=8. Ketiadaan korpus yang komprehensif menghambat penelitian dalam pemrosesan bahasa alami dan aplikasi kecerdasan buatan, yang memerlukan data bahasa yang terstruktur.

https://doi.org/10.25077/TEKNOSI.v11i1.2025.17-26

1.2. Kesalahan Ejaan dalam Bahasa Indonesia

Pada dasarnya kesalahan ejaan dalam bahasa daerah termasuk Bahasa Minang memiliki kesamaan dengan kesalahan ejaan dalam bahasa Indonesia. Kesalahan ejaan dalam Bahasa Indonesia meliputi beberapa bagian yaitu kesalahan fonologi, morfologis, homofon, transliterasi, *typografi*, dan kesalahan penulisan kata serapan. Kesalahan fonologi adalah kesalahan yang timbul karena ejaan yang mendekati bunyi kata. Contoh: *kuda* ditulis sebagai *kudah*. Kesalahan morfologis adalah kesalahan dalam pembentukan kata, khususnya saat menggunakan afiks atau kata dasar. Contoh: *bermain* menjadi *bermainan* (karena pengaruh logika morfologi yang salah) [21]. Kesalahan homofon terkait penggunaan kata yang salah tetapi memiliki bunyi yang sama. Contoh: *masa* (waktu) ditulis sebagai *massa* (kumpulan). Kesalahan transliterasi dipengaruhi oleh penulisan dalam aksara lain, seperti aksara daerah. Contoh: *suku Minangkabau* ditulis *suko Minangkabau* (terpengaruh fonetik lokal). Kesalahan *typografi* merupakan kesalahan teknis seperti penekanan tombol yang salah. Contoh: *tulis* menjadi *tukis*, *bahasa* menjadi *bahasq* dan selanjutnya kesalahan penulisan kata serapan yaitu kata serapan dari bahasa asing ditulis tidak sesuai kaidah ejaan. Contoh: *analisis* menjadi *analisa*, *aktivitas* menjadi *aktifitas* [22]. Kesalahan ejaan lainnya yaitu kesalahan kata nyata yaitu merupakan kesalahan penulisan suatu kata namun kata salah yang dihasilkan adalah merupakan kata lain dengan makna yang berbeda dan digunakan dalam konteks berbeda. Contoh: *bisa* jika salah tulis dapat menjadi *busa* yang maknanya berbeda, *kapur* jika salah tulis dapat menjadi *kasur* dan lain-lain.

1.3. Kesalahan Typografi

Kesalahan ejaan yang disebabkan oleh *typo* (*typographical errors*) adalah kesalahan teknis yang terjadi selama proses pengetikan. Kesalahan ini biasanya bukan akibat kurangnya pemahaman terhadap aturan bahasa, tetapi karena faktor manusia, seperti tergesa-gesa, kurangnya perhatian, atau kesalahan mekanis saat mengetik. Kesalahan yang termasuk dalam kategori *typo* meliputi beberapa bagian yaitu kesalahan penyisipan karakter, penghapusan karakter, penggantian karakter, pertukaran posisi karakter, penghilangan spasi atau tanda baca, penggunaan huruf kapital, dan kesalahan kata homofon.

Kesalahan penyisipan merupakan kesalahan penambahan karakter yang tidak seharusnya ada dalam kata. Biasanya terjadi karena tekanan tombol *keyboard* yang tidak disengaja. Contoh: *kata* menjadi *kaata*, *mobil* menjadi *mobill.* Kesalahan penghapusan karakter yaitu hilangnya karakter yang seharusnya ada, sering kali terjadi karena tombol *keyboard* tidak ditekan dengan cukup kuat. Contoh: *bahasa* menjadi *bhasa*, *pergi* menjadi *prig.* Kesalahan penggantian karakter terjadi dikarenakan karakter yang benar digantikan dengan karakter yang salah, biasanya karena tombol yang ditekan berdekatan pada *keyboard*. Contoh: *tulis* menjadi *tukis*, *besar* menjadi *beser.* Kesalahan pertukaran posisi terjadi ketika dua karakter yang berdekatan tertukar posisinya selama proses pengetikan. Contoh: *rumah* menjadi *rumha*, *taman* menjadi *tamna.* Kesalahan tanda baca, seperti 100% menjadi 100^, *Awas!* menjadi *Awas@*. Penambahan karakter, penggantian karakter, penghapusan karakter, dan pindah posisi karakter pada kata yang *typo* dapat lebih dari satu karakter. Penggunaan huruf besar atau kecil digunakan secara tidak tepat, biasanya terjadi karena pengaturan keyboard atau tekanan tombol shift yang salah. Contoh: *Bahasa* menjadi *bahasa*, *INDONESIA* menjadi *Indonesia.* Kesalahan kata homofon terjadi ketika kata yang diucapkan sama, tetapi ejaannya berbeda, digunakan secara salah. Meskipun ini bisa bersifat kognitif, sering kali terjadi juga karena *typo*. Contoh: *masa* menjadi *massa, teh* menjadi *tah* [23][24]. Tabel 1 menampilkan beberapa contoh kesalahan *typografi*.

Tabel 1. Kesalahan *Typografi*

Jenis kesalahan	**Kata (benar)**	**Kata (*typo*)**
Penyisipan karakter	kata mengikuti pembentukan	kaata mmenngikuti peembentukann
Penghapusan karakter	bahasa khususnya dikelompokkan	bhasa khsusya dkelmpokan
Penggantian karakter	tulis merupakan dipermainkannyalah	tukis neruoakan dipefmainkamnyalag
Pertukaran posisi	rumah komponen pemahamannyalah	rumha kopmonne pmeahaamnnyalha
Kesalahan tanda baca	100% Awas! Di rumah.	100^ Awas@ Di rumah/

1.4. Struktur Bahasa Minang

Struktur atau tata Bahasa Minangkabau dapat dikelompokkan ke dalam beberapa komponen utama, yaitu fonologi, morfologi, dan sintaksis. Tata Bahasa Minangkabau memiliki sistem fonologi yang khas. Misalnya, fonem dalam Bahasa Minangkabau mencakup vokal dan konsonan yang kadang berbeda distribusinya dengan Bahasa Indonesia. Contoh unik adalah variasi bunyi dalam dialek-dialek Minangkabau seperti bunyi nasal yang diintegrasikan kedalam kata-kata tertentu [25].
Urutan kata dalam kalimat Bahasa Minangkabau cenderung mengikuti pola subjek-predikat-objek (SPO). Namun, dalam beberapa konteks, pola ini dapat berubah tergantung pada fokus kalimat. Kalimat interogatif biasanya ditandai dengan partikel atau perubahan intonasi tanpa perubahan struktur dasar kalimat [25].

Morfologi merupakan studi tentang struktur dan pembentukan kata. Dalam Bahasa Minangkabau, pembentukan kata sering melibatkan proses berikut yaitu afiksasi, reduplikasi, pengubahan kata dasar, pemakaian partikel, dan penggunaan kata serapan. Afiksasi atau penambahan afiks terdiri atas prefiks (awalan). Contohnya adalah *ba-* (menunjukkan aktivitas atau keadaan), *ma-* (serupa dengan "ber-" dalam Bahasa Indonesia), dan *paN-* (menyatakan pekerjaan atau pelaku). Misalnya: *bapandai* (pandai berbuat sesuatu), *mancari* (mencari), *pangaluak* (orang yang mengelak atau berkelit). Sufiks (Akhiran) seperti *-an* dan *-nyo* digunakan untuk membentuk kata benda atau mempertegas makna. *Rumah-an* (berkaitan dengan rumah), *inyo* (miliknya).

Konfiks (Gabungan Awalan dan Akhiran) seperti *ba-...-an*, yang sering dipakai untuk menyatakan aktivitas: *batulisan* (sedang menulis) [7]. Reduplikasi (Pengulangan) merupakan proses pengulangan kata yang digunakan untuk menyatakan makna intensitas, kuantitas, atau bentuk kolektif. Contoh: *anak-anak* (banyak anak), *jalan-jalan* (berjalan-jalan atau aktivitas rekreasi).

Kata dasar dapat mengalami modifikasi untuk menunjukkan bentuk baru, misalnya melalui penyisipan atau penghilangan fonem tertentu, seperti: *tangih* menjadi *mangangih* (menangis). Bahasa Minangkabau sering menggunakan partikel untuk memperjelas atau menambah makna kata, seperti partikel *lah* untuk menekankan pernyataan atau kondisi, misalnya: *lai datanglah* (sudah datang). Bahasa Minangkabau juga menyerap kata dari bahasa lain, terutama Bahasa Melayu Kuno, Arab, dan Indonesia, namun sering kali dengan modifikasi fonologi atau morfologi sehingga sesuai dengan struktur Minangkabau. Contohnya: *suro* (asal kata Arab "syuro" berarti musyawarah) [26]

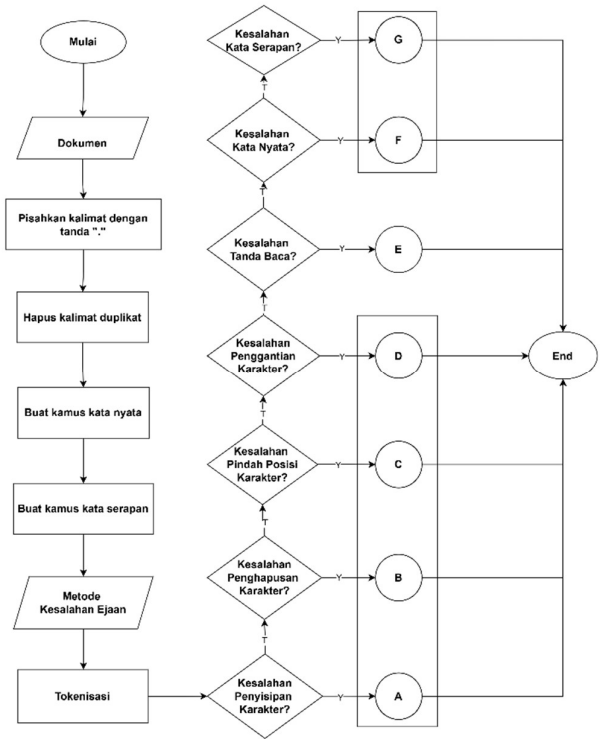
Kesalahan ejaan dalam bahasa Minang dapat dikelompokkan berdasarkan jenis-jenis kesalahan ejaan seperti pada tabel 2. Penggunaan kata serapan dalam Bahasa Minang biasanya akan sama mengikuti kata serapan dalam Bahasa Indonesia.

Tabel 2. Kesalahan ejaan pada Bahasa Minang

Jenis kesalahan	**Kata Minang (benar)**	**Kata Minang (*typo*)**
	pai	paii
Penyisipan karakter	mangikuik dilantakkannyo	maangikuikk Dillantaakkannyoo
Penghapusan karakter	bahaso ditaragak sadoalahnyo	bhaso dtragak sdoalhny
Penggantian karakter	balaja saumpamonyo manggunokannyopun	bilaja saunpanonyo nanggumokannyopum
Pertukaran posisi	kasua bamainan manunjuakkan	sakua bmaainna namunjukakna
Kesalahan tanda baca	100% Awas! di rumah.	100^ Awas@ di rumah/
Kesalahan kata nyata	kambuah pai namo	Tambuah lai lamo
Kesalahan kata serapan	kuitansi Praktik aktivitas	kwitansi praktek aktifitas

2. METODE

Metode pada penelitian ini secara garis besar melalui dua tahapan utama, yaitu pengumpulan data Bahasa Minang dan pembentukan korpus kesalahaan ejaan dalam Bahasa Minang. Gambar 1 menampilkan metodologi penelitian.



Gambar 1. Metodologi Penelitian

 https://doi.org/10.25077/TEKNOSI.v11i1.2025.17-26

2.1. Pengumpulan Data

Pengumpulan data kalimat Bahasa Minang didapat dari beberapa sumber, yaitu *wikipedia dump*, portal berita, lirik lagu, dan pepatah. Kalimat Bahasa Minang bervariasi dari jumlah token dalam satu kalimat. Kalimat pendek terdiri atas minimal tiga token dalam satu kalimat, sedangkan kalimat panjang dapat terdiri sampai dengan 30 token dalam satu kalimat.

2.2. Pembentukan Korpus

Tahapan pembentukan korpus dapat dilihat pada Gambar 1. Proses dimulai dengan menyiapkan data kalimat Bahasa Minang. Proses pemisahan setiap kalimat pada dokumen dengan tanda ".". Tujuan dari proses pemisahan ini agar teks tidak terlalu panjang dan menambah *dataset*. Contoh: "Dek itu, mereka mandukuang jikok pemerintah mambuek aturan yang labiah ketat dalam pamakaian media sosial. Urang Amerika Serikat (AS) alah mulai berang jo banyaknyo hoaks atau kaba duto, nan baredar di dunia maya". Teks tersebut dapat dipisahkan dengan tanda "." menjadi dua kalimat sebagai berikut: 1) "Dek itu, mereka mandukuang jikok pemerintah mambuek aturan yang labiah ketat dalam pamakaian media sosial.", 2) "Urang Amerika Serikat (AS) alah mulai berang jo banyaknyo hoaks atau kaba duto, nan baredar di dunia maya." Pemisahan setiap kalimat dilakukan agar tidak banyak data yang terbuang, karena pada proses pembuatan korpus kesalahan ejaan hanya satu kata/*token* yang akan dilakukan proses pengacakan. Kalimat ganda atau kalimat yang sama harus dihapus dari dataset, dikarenakan kalimat yang diproses harus unik dan akan memperbesar jumlah *dataset*.

Kesalahan kata nyata merupakan proses token diganti dengan token lain yang berbeda konteks. Pembuatan kamus kata nyata dilakukan dengan cara otomatis menggunakan algoritma *Damerau Levenstein Distance* (DLD). Algoritma ini menghitung jumlah operasi minimum *edit* untuk mengubah satu string kedalam bentuk string lainnya, atau dengan kata lain mencari kemiripan antar string kata. Operasi yang dilakukan adalah sebagai berikut: *insertion* (penyisipan), *deletion* (penghapusan), *subtitution* (pergantian), dan *transposition* (pertukaran). Kamus kata nyata memiliki format *dictionary* dengan cara membuat "kata asli":"kata pengganti". Proses akan mengubah setiap kata yang terdaftar sebagai "kata asli" menjadi "kata pengganti". Contoh : kata *pasa* setelah ditemukan dalam kamus maka akan diganti dengan kata *dasa*. *Pasa* artinya pasar dan *dasa* artinya dasar. Berdasarkan rumus algoritma DLD maka dapat ditentukan jarak perbedaan antara dua kalimat tersebut. Detail dari algoritma DLD tidak dibahas dalam penelitian ini. Beberapa contoh lainnya untuk kata nyata dalam bahasa Minang dapat dilihat pada Tabel 3.

Tabel 3. Kamus Kata Nyata

Kata awal (arti)	Kata	Jarak Kemiripan
Pasa (pasar)	Dasa (dasar)	1
Pai (pergi)	lai (iya)	1
Tambuah (tambah)	Kambuah (kambuh)	1
Baserak (berserakan)	Basurak (bersorak)	1

Kamus kata serapan dibuat untuk mencocokkan kata yang tertulis dengan kata serapan tersebut dengan ejaan yang sering salah dilakukan. Tabel 4 menunjukkan beberapa daftar kata serapan dan kesalahan yang sering terjadi. Contoh: kata *kuitansi* adalah penulisan yang benar dan *kwitansi* adalah penulisan yang salah.

Tabel 4. Kamus Kata Serapan

Kata serapan (benar)	Kata serapan (salah)
Kuitansi	Kwitansi
Ijazah	Ijasah
Analisis	Analisa
Praktik	Praktek

Metode kesalahan ejaan adalah metode yang digunakan untuk menghasilkan dokumen/kalimat yang terdapat kesalahan ejaan di dalamnya. Terdapat tujuh metode kesalahan ejaan yang dikerjakan dalam penelitian ini, yaitu: 1) kesalahan penyisipan karakter, 2) kesalahan penghapusan karakter, 3) kesalahan pindah posisi karakter, 4) kesalahan penggantian karakter, 5) kesalahan tanda baca, 6) kesalahan kata nyata, dan 7) kesalahan kata serapan. Setiap kali proses hanya dapat memilih salah satu metode agar dapat menghasilkan satu *dataset* baru hasil dari metode yang dipilih.

Tokenisasi merupakan proses memotong *token* pada kumpulan kalimat/dokumen. Tokenisasi dilakukan agar lebih mudah memilih token yang akan dilakukan pengacakan pada karakternya. Sistem akan bekerja melakukan pengacakan karakter berdasarkan inputan metode kesalahan ejaan. Dari tujuh kesalahan ejaan yang dilakukan dikelompokkan menjadi tiga kategori berdasarkan tahapan pembentukan korpusnya yaitu kelompok pertama terdiri atas kesalahan penyisipan karakter (A), kesalahan penghapusan karakter (B), kesalahan pindah posisi karakter (C), dan kesalahan penggantian karakter (D). Kelompok kedua yaitu kesalahan tanda baca (E) dan kelompok ketiga yaitu kesalahan kata nyata (F) dan kesalahan kata serapan (G).

Gambar 2 menampilkan proses pembentukan korpus kesalahan penyisipan karakter sebagai lanjutan dari gambar 1 jika inputan metode kesalahan ejaannya adalah penyisipan karakter. Sistem akan melakukan proses tokenisasi dari semua kalimat dalam data set dan akan menghasilkan daftar token dari setiap kalimat. Selanjutnya sistem menghitung semua panjang token (L) dan mengelompokkan semua token berdasarkan panjang token. Dalam hal ini terdapat tiga pengelompokan berdasarkan jumlah karakter yang akan diacak pada token, yaitu seperti yang dinyatakan pada tabel 5.

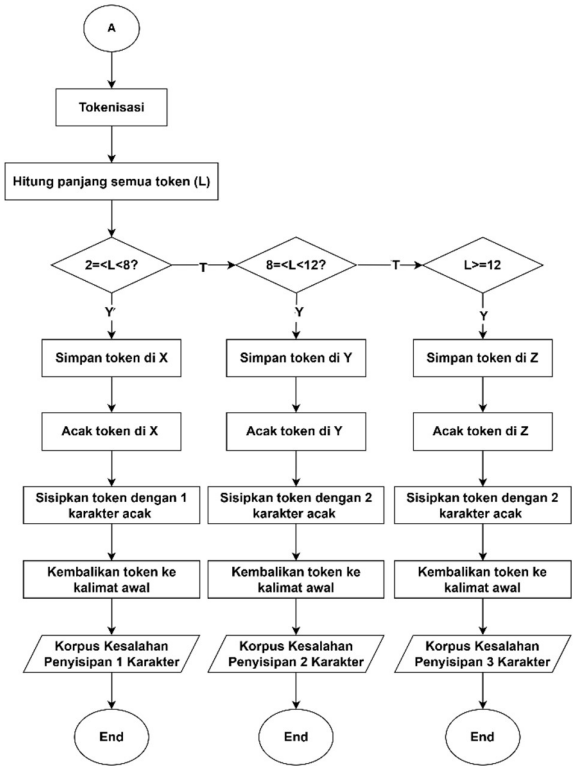
Tabel 5. Pengelompokan Token dan Jumlah Karakter Acak

Panjang Token (L)	Jml Karakter diacak	Contoh Token
2=<L<8	1	Ibo
8=<L<12	2	Mambuang
L>=12	3	pabandiangan

Token yang telah dikelompokkan disimpan dalam tiap variabel untuk selanjutnya dilakukan pengacakan token pada tiap kelompok token berdasarkan panjang token. Token kelompok pertama 2=<L<8 akan disisipkan 1 karakter acak, token

kelompok kedua 8=<L<12 akan disisipkan 2 karakter acak dan token L>=12 akan disisipkan 3 karakter acak. Proses dilanjutkan dengan mengembalikan token ke kalimat awal dan terakhir dihasilkan daftar korpus kesalahan penyisipan sebanyak 1 karakter, 2 karakter dan 3 karakter. Proses ini sama untuk tiga metode kesalahan ejaan lainnya yaitu kesalahan penghapusan karakter (B), kesalahan pindah posisi karakter (C) dan kesalahan penggantian karakter (C). Tahap akhir didapat daftar kelompok korpus untuk tiap metode kesalahan ejaan dan pengacakan sampai dengan tiga karakter.

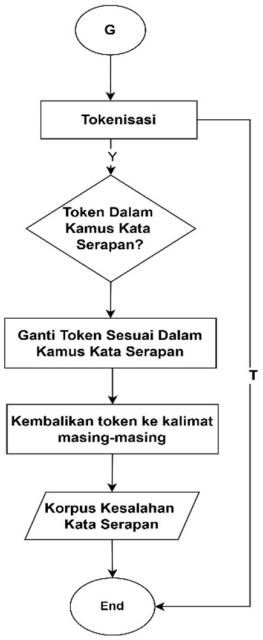
Gambar 3 menampilkan proses pembentukan korpus kesalahan tanda baca. Sistem melakukan proses tokenisasi dari semua kalimat dalam data set dan akan menghasilkan daftar token dari setiap kalimat. Selajutnya token yang memiliki tanda baca akan diacak dan selanjutnya ganti tanda baca dengan tanda baca lainnya dan kembalikan token ke kalimat masing-masing.



Gambar 2. Pembentukan Korpus Kesalahan Penyisipan Karakter

Gambar 3. Pembentukan Korpus Kesalahan Tanda Baca

Gambar 4 menampilkan proses pembentukan kesalahan kata serapan. Sistem melakukan proses tokenisasi dari semua kalimat dalam dataset dan akan menghasilkan daftar token dari setiap kalimat. Token yang didapat dari tokenisasi dicek dalam kamus kata serapan. Hal ini dilakukan untuk membandingkan apakah token tersebut termasuk kata serapan atau tidak. Jika token ada pada kamus kata serapan maka ganti token tersebut dengan token padanannya dalam kamus tersebut selanjutnya kembalikan token ke kalimat awal dan terakhir dihasilkan korpus kesalahan kata serapan. Jika token tidak terdapat pada kamus serapan maka proses berhenti.



Gambar 4. Kesalahan Kata Serapan.

Proses pada pembentukan korpus kesalahan kata serapan sama halnya dengan pembentukan korpus kata nyata hanya terdapat perbedaan di kamus sumbernya.

 https://doi.org/10.25077/TEKNOSI.v11i1.2025.17-26

3. HASIL DAN PEMBAHASAN

Proses pembentukan korpus kesalahan ejaan yang dikembangkan meliputi tujuh kesalahan.

3.1. Kesalahan Penyisipan/Penghapusan/Pindah Posisi/Penggantian Karakter

Pada pembentukan korpus ini dilakukan penyisipan karakter, penghapusan karakter, pindah posisi karakter dan penggantian karakter sejumlah satu karakter, dua karakter dan tiga karakter secara acak pada kelompok token berdasarkan panjang token seperti yang dinyatakan pada tabel 5. Contoh : *indonesia* menjadi *inndonesiaa*, terdapat penyisipan 2 karakter yaitu 'n' dan 'a'. Pseudocode 1 menjelaskan langkah untuk menghasilkan korpus kesalahan penyisipan karakter sampai dengan tiga karakter yang disisipkan pada token. Hal ini berlaku untuk tiga kesalahan ejaan lainnya.

	Pseudocode 1 Kesalahan Penyisipan Karakter
1	**Input :** Dataset, Metode kesalahan ejaan
2	**Output:** Korpus kesalahan penyisipan satu karakter, *k*orpus kesalahan penyisipan dua karakter, korpus
3	kesalahan penyisipan tiga karakter
4	
5	**Begin**
6	**For** setiap kalimat dalam dataset **do**
7	Tokenisasi(kalimat)
8	Hitung panjang semua token (L)
9	**If** 2=<L<8
10	**then** Simpan token ke X
	Acak token dalam X
11	Sisipkan **satu** karakter dalam token
12	Kembalikan token ke kalimat semula (X1)
13	**Else if** 8=<L<12
14	**then** Simpan token ke Y
15	Acak token dalam Y
16	Sisipkan **dua** karakter dalam token
17	Kembalikan token ke kalimat semula
18	(X2)
19	**Else if** L>=12
	then Simpan token ke Z
	Acak token dalam Z
	Sisipkan tiga karakter
	Kembalikan token ke kalimat semula (X3)
	End if
	End if
	End if
	End

Implementasi dari pseudocode 1 dapat dilihat pada Tabel 6. Contoh dokumen inputnya adalah kalimat "pangkeknyo dinaiakkan dari brigadir manjadi mayor jenderal". Dokumen dilakukan tokenisasi kemudian dihitung panjang tiap token. Selanjutnya token diacak dan diambil satu token untuk ditambahkan sejumlah karakter. Kata yang ditandai tebal adalah kata yang didapat dari hasil pengacakan sekelompok token.

Tabel 6. Pengacakan Kesalahan Penyisipan Karakter

Daftar Token	L	X	Y	Z
pangkeknyo	10			pangkeknyo
dinaiakkan	9			**dinaiakkan**
dari	4	**dari**		
brigadir	8		brigadir	
manjadi	7	manjadi		
mayor	5	mayor		
jenderal	8		**jenderal**	

Ket:
L adalah panjang token
X adalah kelompok token dengan 2=<L<8
Y adalah kelompok token dengan 8=<L<12
Z adalah kelompok token dengan L>=12

Proses ini berlaku sama untuk tiga kesalahan ejaan lainnya. Tabel 7 menampilkan hasil pengembangan korpus pada empat kesalahan ejaan dengan pengacakan sampai dengan tiga karakter.

Tabel 7. Hasil pengacakan kesalahan penyisipan karakter

Metode	Kalimat hasil kesalahan penyisipan karakter
X1	pangkeknyo dinaiakan **darii** brigadir manjadi mayor jenderal
X2	pangkeknyo dinaiakan dari brigadir manjadi mayor **jeenderall**
X3	pangkeknyo **dinnaiaakkann** dari brigadir manjadi mayor jenderal
Metode	**Kalimat hasil kesalahan penghapusan karakter**
X1	pangkeknyo dinaiakan **dri** brigadir manjadi mayor jenderal
X2	pangkeknyo dinaiakkan dari brigadir manjadi mayor **jndral**
X3	pangkeknyo **dnaikan** dari brigadir manjadi mayor jenderal
Metode	**Kalimat hasil kesalahan pindah posisi karakter**
X1	pangkeknyo dinaiakan **drai** brigadir manjadi mayor jenderal
X2	pangkeknyo dinaiakkan dari brigadir manjadi mayor **jnederla**
X3	pangkeknyo **idnaaikkna** dari brigadir manjadi mayor jenderal
Metode	**Kalimat hasil kesalahan penggantian karakter**
X1	pangkeknyo dinaiakkan **dsri** brigadir manjadi mayor jenderal
X2	pangkeknyo dinaiakkan dari brigadir manjadi mayor **jemdsral**
X3	pangkeknyo **simaiaklan** dari brigadir manjadi mayor jenderal

Ket:
X1 adalah kalimat hasil pengacakan 1 karakter
X2 adalah kalimat hasil pengacakan 2 karakter
X3 adalah kalimat hasil pengacakan 3 karakter

3.2. Kesalahan Tanda Baca

Kesalahan tanda baca merupakan proses mengganti tanda baca dalam kalimat dengan tanda baca lainnya secara acak. Contoh:

laki-laki menjadi laki+laki. Pseudocode 2 menjelaskan langkah untuk menghasilkan kalimat dengan salah satu tokennya memiliki kesalahan tanda baca.

	Pseudocode 2 Tanda Baca
1	**Input :** Dataset, Metode kesalahan ejaan
2	**Output:** Korpus kesalahan tanda baca
	Begin
3	**For** setiap kalimat dalam dataset **do**
4	Tokenisasi(kalimat)
5	Cari token dengan tanda baca
6	Acak token dengan tanda baca
7	Ganti tanda baca dengan tanda baca lainnya
8	Kembalikan token ke kalimat semula
9	**End**

Implementasi dari pseudocode 2 dapat dilihat pada Tabel 8. Contoh: "laki-laki 54% dari jumlah populasi dan padusi 46%."

Tabel 8. Hasil Kesalahan Tanda Baca

Token	**Token dengan tanda baca**	**Token yang diacak**
laki-laki	**laki-laki**	**Laki+laki**
54%	54%	
dari		
jumlah		
populasi		
kasadonyo		
dan		
padusi		
46%.	46%.	

Kalimat yang dihasilkan dari proses kesalahan tanda baca yaitu : "**laki+laki** 54% dari jumlah populasi dan padusi 46%."

3.3. Kesalahan Kata Serapan

Kesalahan kata serapan merupakan kesalahan ejaan dikarenakan ketidaktahuan penulisan yang benar dari kata serapan. Contoh diberikan kalimat "inyo mamagang kwitansi pambalian", kemudian dilakukan tokenisasi. Semua daftar token akan dicari dalam kamus kata serapan. Jika token ada maka ganti token dengan token dalam kamus, jika token tidak ada maka proses selesai. Pseudocode 3 menjelaskan langkah untuk menghasilkan kalimat dengan kesalahan kata serapan.

	Pseudocode 3 Kesalahan Kata Serapan
1	**Input :** Dataset, Metode kesalahan ejaan
2	**Output:** Korpus kesalahan kata serapan
3	**Begin**
4	**For** setiap kalimat dalam dataset **do**
5	Tokenisasi(kalimat)
6	**If** token dalam kamus kata serapan
7	**then** ganti token sesuai kamus
8	**End if**
9	Kembalikan token kalimat semula
10	**End**

Implementasi dari pseudocode 3 dapat dilihat pada Tabel 9. Contoh: "inyo mamagang kwitansi pambalian".

Tabel 9. Hasil Kesalahan Kata Serapan

Token	**Token di kamus**	**Kesalahan kata serapan**
inyo		
mamagang		
kwitansi	kwitansi	kuitansi
pambalian		

Kalimat yang dihasilkan dari proses tersebut yaitu : "inyo mamagang **kwitansi** pambalian".
Pembentukan korpus kesalahan kata nyata memiliki alur yang sama dengan kesalahan kata serapan seperti diyatakan pada tabel 10. Kalimat yang dihasilkan yaitu : "adiak **lai** mambali obat maghnyo **tambuah**"

Tabel 10. Hasil Kesalahan Kata Nyata

Token	**Token di kamus**	**Kata lain**
adiak		
Pai	**pai**	**lai**
mambali		
obat		
Maghnyo		
kambuah	**kambuah**	**tambuah**

Dari tujuh proses pembentukan kesalahan ejaan yang dilakukan dihasilkan 15 file kumpulan korpus seperti yang dinyatakan pada tabel 10.

Tabel 10. Daftar Korpus Kesalahan Ejaan Bahasa Minang

No	**Jenis Korpus**
1.	Kesalahan penyisipan 1 karakter
2.	Kesalahan penyisipan 2 karakter
3.	Kesalahan penyisipan 3 karakter
4.	Kesalahan penghapusan 1 karakter
5.	Kesalahan penghapusan 2 karakter
6.	Kesalahan penghapusan 3 karakter
7.	Kesalahan pindah posisi 1 karakter
8.	Kesalahan pindah posisi 2 karakter
9.	Kesalahan pindah posisi 3 karakter
10.	Kesalahan penggantian 1 karakter
11.	Kesalahan penggantian 2 karakter
12.	Kesalahan penggantian 3 karakter
13.	Kesalahan tanda baca
14.	Kesalahan kata nyata
15.	Kesalahan kata serapan

https://doi.org/10.25077/TEKNOSI.v11i1.2025.17-26

4. KESIMPULAN

Penelitian ini menghasilkan SPEML yang merupakan korpus kesalahan ejaan dalam Bahasa Minang yang terdiri atas tujuh kesalahan yaitu kesalahan pengacakan sampai dengan tiga karakter pada kesalahan penyisipan karakter, kesalahan penghapusan karakter, kesalahan pindah posisi karakter dan kesalahan penggantian karakter, kesalahan tanda baca diikuti dengan kesalahan kata nyata, dan kesalahan kata serapan. Algoritma SPEML yang dihasilkan dapat diterapkan untuk semua bahasa daerah lainnya. Pengembangan korpus SPEML merupakan langkah penting untuk memungkinkan pelatihan dan evaluasi model NLP yang pada akhirnya meningkatkan akurasi dalam mengidentifikasi dan memperbaiki kesalahan ejaan dalam teks Bahasa Minang.

DAFTAR PUSTAKA

[1] G. Indra Winata *et al.*, "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages." [Online]. Available: https://github.com/

[2] K. P. dan K. Badan Pengembangan Bahasa dan Perbukuan, "Bahasa dan Peta Bahasa di Indonesia," https://petabahasa.kemdikbud.go.id/index.php.

[3] S. Raharjo, E. Utami, M. Yusa, and E. Sutanta, "Systematic Literature Review: Corpus Linguistics in Indonesia," in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 370–377. doi: 10.1007/978-3-031-06417-3_50.

[4] J. A. Lopo and R. Tanone, "Constructing and Expanding Low-Resource and Underrepresented Parallel Datasets for Indonesian Local Languages," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.01009

[5] W. Wongso, A. Joyoadikusumo, B. S. Buana, and D. Suhartono, "Many-to-Many Multilingual Translation Model for Languages of Indonesia," *IEEE Access*, vol. 11, pp. 91385–91397, 2023, doi: 10.1109/ACCESS.2023.3308818.

[6] R. Sovia, S. Defit, and Yuhandri, "Development of the Minangkabau Local Language Translation Machine Based on Stemming," in *Proceeding - 2022 International Symposium on Information Technology and Digital Innovation: Technology Innovation During Pandemic, ISITDI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 195–198. doi: 10.1109/ISITDI55734.2022.9944457.

[7] A. Ayub *et al.*, "Tata Bahasa Minangkabau," p. 234, 1993.

[8] Y. M. Altameemi, "State-of-the-Art Review of the Corpus Linguistics Field From the Beginning Until the Development of ChatGPT," *Theory and Practice in Language Studies*, vol. 14, no. 2, pp. 423–431, Feb. 2024, doi: 10.17507/tpls.1402.13.

[9] J. A. Lopo and R. Tanone, "Constructing and Expanding Low-Resource and Underrepresented Parallel Datasets for Indonesian Local Languages," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.01009

[10] Y. Yanfi, F. L. Gaol, B. Soewito, and H. L. H. S. Warnars, "Spell Checker for the Indonesian Language: ExtensiveReview," *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, no. 5, pp. 1–7, May 2022, doi: 10.46338/ijetae0522_01.

[11] D. A. Kwary, "A corpus platform of Indonesian academic language," *SoftwareX*, vol. 9, pp. 102–106, Jan. 2019, doi: 10.1016/j.softx.2019.01.011.

[12] S. Raharjo, E. Utami, M. Yusa, and E. Sutanta, "Systematic Literature Review: Corpus Linguistics in Indonesia," in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 370–377. doi: 10.1007/978-3-031-06417-3_50.

[13] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji, "Cross-lingual name tagging and linking for 282 languages," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Association for Computational Linguistics (ACL), 2017, pp. 1946–1958. doi: 10.18653/v1/P17-1178.

[14] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," Aug. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2023.e18647.

[15] F. Koto and I. Koto, "Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation." [Online]. Available: https://id.wikimedia.org/wiki/

[16] O. V. Putra, F. M. Wasmanson, T. Harmini, and S. N. Utama, "Sundanese Twitter Dataset for Emotion Classification," in *CENIM 2020 - Proceeding: International Conference on Computer Engineering, Network, and Intelligent Multimedia 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 391–395. doi: 10.1109/CENIM51130.2020.9297929.

[17] N. Ifada, F. H. Rachman, M. W. M. A. Syauqy, S. Wahyuni, and A. Pawitra, "MadureseSet: Madurese-Indonesian Dataset," *Data Brief*, vol. 48, Jun. 2023, doi: 10.1016/j.dib.2023.109035.

[18] A. Mohammed Saleh Al-Hamzi, A. Gougui, Y. Sari Amalia, and T. Suhardijanto, "Corpus Linguistics and Corpus-Based Research and its Implication in Applied Linguistics: A Systematic Review," *Parole: Journal of Linguistics and Education*, vol. 10, no. 2, pp. 2020–176, 2020.

[19] Y. Yanfi, R. Setiawan, H. Soeparno, and W. Budiharto, "SPECIL: Spell Error Corpus for the Indonesian Language," *IEEE Access*, vol. 11, pp. 93227–93237, 2023, doi: 10.1109/ACCESS.2023.3307712.

[20] Y. Yanfi, R. Setiawan, H. Soeparno, and W. Budiharto, "Comparison of Spelling Error Correction Algorithms for the Indonesian Language," in *2023 11th International Conference on Information and Education Technology, ICIET 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 443–447. doi: 10.1109/ICIET56899.2023.10111191.

[21] "2021-ACM-A Framework for Indonesian Grammar Error Correction".

[22] D. A. Anggoro and I. Nurfadilah, "Active Verb Spell Checking Mem- + P in Indonesian Language Using the

Jaro-Winkler Distance Algorithm," *Iraqi Journal of Science*, vol. 63, no. 4, pp. 1811–1822, 2022, doi: 10.24996/ijs.2022.63.4.38.

[23] "Pembangunan Aplikasi Identifikasi Kesalahan Ketik Jaro Winkler Distance".

[24] A. Amalia, O. S. Sitompul, T. Mantoro, and E. B. Nababan, "Morpheme Embedding for Bahasa Indonesia Using Modified Byte Pair Encoding," *IEEE Access*, vol. 9, pp. 155699–155710, 2021, doi: 10.1109/ACCESS.2021.3128439.

[25] F. Rahman, S. Kurniati, and Nova Rina, "Basis Data Leksikal: Perubahan Bunyi Bahasa Minangkabau Isolek Sangir Jujuan," *Linguistik Indonesia*, vol. 42, no. 1, pp. 185–198, 2024, doi: 10.26499/li.v42i1.572.

[26] H. Priyatman, M. Saleh, and H. Sujaini, "Analisis Akurasi Algoritma Extended Word Similarity Based Clustering (EWSB) pada Mesin Penerjemah Bahasa Indonesia-Minang," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 6, no. 3, p. 323, 2020, doi: 10.26418/jp.v6i3.43330.

BIODATA PENULIS

Dewi Soyusiawaty adalah dosen di Program Studi Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan. Minat penelitiannya meliputi Pemrosesan Bahasa Alami, Sistem Penerjemahan, dan Kesalahan Ejaan.

Abdul Fadlil adalah Guru Besar di Program Studi Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan. Bidang keahliannya meliputi Elektronika, Sistem Cerdas, dan Pemrosesan Sinyal.

Sunardi adalah Guru Besar di Program Studi Teknik Elektro, Fakultas Teknologi Industri, Universitas Ahmad Dahlan. Bidang keahliannya meliputi Teori Informasi, Sistem Komunikasi, dan Komunikasi Data.

-