

Pemanfaatan Skor Ejaan Berbasis Levenshtein Distance sebagai Fitur Klasifikasi Risiko Literasi Siswa SD

^{1*}Ciko Nanda Putra Wardana, ²Intan Nur Farida, ³Danar Putra Pamungkas

^{1,2,3} Teknik Informatika, Universitas Nusantara PGRI Kediri

E-mail: ¹cikonanda08@gmail.com, ²intannf@unpkediri.ac.id, ³danar@unpkediri.ac.id

Penulis Korespondens : Ciko Nanda Putra Wardana

Abstrak— Kemampuan literasi pada siswa sekolah dasar dipengaruhi oleh berbagai aspek, di antaranya kemampuan membaca, menulis, dan memahami materi pembelajaran. Ketepatan ejaan merupakan salah satu indikator yang dapat digunakan untuk menilai kemampuan menulis siswa. Penelitian ini bertujuan mengembangkan pendekatan deteksi risiko literasi dengan memanfaatkan skor ejaan hasil analisis Levenshtein Distance sebagai fitur klasifikasi. Metode Levenshtein Distance digunakan untuk mengidentifikasi kesalahan ejaan dan menghasilkan skor ejaan yang selanjutnya dikombinasikan dengan nilai membaca dan nilai kuis. Ketiga fitur tersebut digunakan sebagai masukan pada algoritma Random Forest untuk mengklasifikasikan tingkat risiko literasi siswa. Hasil pengujian menunjukkan bahwa model Random Forest memperoleh nilai akurasi sebesar 86,67%. Analisis feature importance menunjukkan bahwa skor ejaan memberikan kontribusi sebesar 27,89% dalam proses klasifikasi. Temuan ini menunjukkan bahwa skor ejaan dapat dimanfaatkan sebagai indikator pendukung dalam identifikasi risiko literasi siswa sekolah dasar secara lebih objektif dan terukur.

Kata Kunci— Levenshtein Distance, machine learning, Random Forest, risiko literasi, skor ejaan.

Abstract— Elementary school students' literacy skills can be influenced by their reading, writing, and learning comprehension abilities. One indicator that can be used to assess writing ability is spelling accuracy. This study aims to utilize a spelling score based on the Levenshtein Distance method as a feature for literacy risk classification among elementary school students. The Levenshtein Distance method was employed to detect spelling errors and generate spelling scores, which were then combined with reading and quiz scores as classification features using the Random Forest algorithm. The results showed that the Random Forest model was able to classify literacy risk with an accuracy of 86.67%. Feature importance analysis revealed that the spelling score contributed 27.89% to the classification process. These findings indicate that the spelling score can be used as an additional indicator for identifying students' literacy risk. This study is expected to support the early detection of literacy risk among elementary school students.

Keywords— Levenshtein Distance, literacy risk classification, machine learning, Random Forest, spelling score

This is an open access article under the CC BY-SA License.



I. PENDAHULUAN

Kemampuan literasi dasar menjadi salah satu faktor penting dalam proses pembelajaran siswa sekolah dasar karena berkaitan dengan kemampuan memahami informasi, berpikir kritis, dan mendukung keberhasilan akademik pada jenjang pendidikan berikutnya [1]. Literasi dasar tidak

hanya mencakup kemampuan membaca, tetapi juga kemampuan menulis, memahami kosakata, serta menyusun kalimat secara tepat. Rendahnya kemampuan literasi dapat menyebabkan siswa mengalami kesulitan dalam memahami materi pelajaran dan berdampak pada penurunan hasil belajar [2]. Oleh sebab itu, diperlukan upaya deteksi dini untuk mengidentifikasi risiko gangguan literasi agar penanganan dapat dilakukan secara lebih cepat dan tepat sasaran.

Permasalahan literasi dasar masih menjadi tantangan di berbagai negara, termasuk Indonesia. Sebagian siswa sekolah dasar masih mengalami kesulitan membaca, memahami teks, dan menulis sesuai kaidah ejaan Bahasa Indonesia. Kesalahan ejaan dapat dijadikan salah satu indikator kemampuan literasi karena berkaitan dengan kemampuan mengenali huruf, memahami kata, dan membentuk kalimat. Selain itu, rendahnya nilai membaca dan hasil evaluasi pembelajaran juga dapat menunjukkan adanya risiko gangguan literasi pada siswa.

Perkembangan teknologi kecerdasan buatan dan Machine Learning memberikan peluang dalam pengembangan sistem deteksi dini gangguan literasi secara otomatis. Salah satu pendekatan yang dapat diterapkan adalah Natural Language Processing (NLP) untuk menganalisis kesalahan ejaan pada teks tulisan siswa [3]. Dalam penelitian ini, metode Levenshtein Distance digunakan untuk menghitung tingkat kemiripan antara kata yang ditulis siswa dan kata baku pada kamus acuan, sehingga jumlah dan tingkat kesalahan ejaan dapat diidentifikasi secara otomatis.

Meskipun metode Levenshtein Distance telah banyak terbukti efektif untuk mendeteksi dan mengoreksi kesalahan ejaan, sebagian besar penelitian masih berfokus pada perbaikan kata dan pemberian rekomendasi ejaan. Padahal, tingkat kesalahan ejaan yang direpresentasikan melalui skor ejaan berpotensi digunakan sebagai indikator kemampuan literasi siswa. Oleh karena itu, penelitian ini memanfaatkan skor ejaan hasil analisis Levenshtein Distance yang dikombinasikan dengan nilai membaca dan nilai kuis untuk mengklasifikasikan risiko literasi menggunakan algoritma Random Forest, yang dipilih karena mampu menghasilkan performa klasifikasi yang stabil serta memiliki risiko *overfitting* yang rendah.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan sistem deteksi dini risiko gangguan literasi dasar berbasis NLP dan Machine Learning pada siswa sekolah dasar. Sistem memanfaatkan metode Levenshtein Distance untuk menghitung kesalahan ejaan, kemudian menggunakan fitur skor ejaan, jumlah kesalahan, rata-rata jarak edit, nilai membaca, dan nilai kuis sebagai masukan pada model Random Forest untuk mengklasifikasikan tingkat risiko literasi ke dalam kategori *low risk*, *moderate risk*, dan *high risk*. Penelitian ini diharapkan dapat membantu guru dalam melakukan identifikasi awal terhadap siswa yang berisiko mengalami gangguan literasi sehingga intervensi pembelajaran dapat dilakukan lebih cepat dan tepat.

II. METODE

Penelitian ini diawali dengan pengumpulan data siswa yang meliputi teks tulisan, nilai membaca, dan nilai kuis. Teks tulisan kemudian melalui tahap *preprocessing* sebelum dianalisis menggunakan metode Levenshtein Distance untuk mendeteksi kesalahan ejaan dan menghasilkan skor ejaan. Skor tersebut selanjutnya dikombinasikan dengan nilai membaca dan nilai kuis sebagai fitur dalam klasifikasi risiko literasi menggunakan algoritma Random Forest. Evaluasi model dilakukan menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*.

A. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini diperoleh secara langsung dari kegiatan pembelajaran siswa sekolah dasar. Data yang dikumpulkan meliputi teks tulisan siswa, nilai membaca, dan nilai kuis. Teks tulisan berasal dari tugas menulis yang dikerjakan siswa selama proses pembelajaran, sedangkan nilai membaca dan nilai kuis diperoleh dari hasil penilaian yang dilakukan oleh guru.

Teks tulisan digunakan sebagai sumber data utama dalam analisis kesalahan ejaan menggunakan metode Levenshtein Distance. Proses tersebut menghasilkan skor ejaan yang kemudian dikombinasikan dengan nilai membaca dan nilai kuis sebagai fitur dalam klasifikasi risiko literasi. Pemilihan ketiga variabel tersebut didasarkan pada penelitian terdahulu yang menunjukkan bahwa kemampuan membaca dan menulis merupakan komponen utama literasi dasar serta dapat digunakan sebagai indikator dalam mengidentifikasi risiko gangguan literasi pada siswa sekolah dasar.

B. Preprocessing Teks

Tahap *preprocessing* dilakukan untuk menyiapkan data teks agar berada dalam kondisi yang bersih, konsisten, dan siap digunakan pada proses analisis kesalahan ejaan:

1. *Case Folding*, yaitu mengubah seluruh teks tulisan siswa menjadi huruf kecil untuk menghindari perbedaan interpretasi akibat penggunaan kapital yang tidak konsisten.
2. *Tokenisasi*, yaitu Memecah teks menjadi satuan kata agar setiap kata dapat dianalisis secara independen pada tahap perhitungan kesalahan ejaan menggunakan metode Levenshtein Distance.

C. Analisis Kesalahan Ejaan Menggunakan Levenshtein Distance

Algoritma Levenshtein Distance menghitung jarak edit minimum antara kedua kata tersebut berdasarkan operasi penyisipan, penghapusan, dan penggantian karakter. Proses perhitungan dilakukan menggunakan matriks jarak, di mana nilai pada setiap sel merepresentasikan jumlah minimum operasi edit yang diperlukan untuk mengubah sebagian string sumber menjadi string target [4]. Berikut ini rumus algoritma Levenshtein Distance:

$$lev_{a,b}(i,j) = \min lev_{a,b}(i-1,j) + 1 \text{ (Penghapusan)} \quad (1)$$

$$lev_{a,b}(i,j) = \min lev_{a,b}(i,j-1) + 1 \text{ (Penyisipan)} \quad (2)$$

$$lev_{a,b}(i,j) = \min lev_{a,b}(i-1,j-1) + 1, a_i \neq b_j \text{ (Substitusi)} \quad (3)$$

Dimana a adalah string sumber, b adalah string target, i adalah index baris string sumber, j adalah index baris string target, ($a_i \neq b_j$) perlu menambahkan 1(+1) [5].

Dalam penerapannya pada koreksi ejaan, *Levenshtein Distance* digunakan untuk mendeteksi kesalahan tulis dengan membandingkan kata tidak baku terhadap seluruh entri dalam kamus. Semakin kecil jaraknya, semakin besar kemungkinan kata tersebut merupakan kandidat kata baku yang relevan [5].

D. Perhitungan Skor Ejaan

Setelah proses analisis kesalahan ejaan selesai, sistem menghitung skor ejaan untuk setiap teks siswa. Skor ini digunakan untuk menggambarkan tingkat ketepatan penulisan dengan membandingkan jumlah kata yang benar terhadap total kata yang dianalisis. Perhitungan skor ejaan mengacu pada penyederhanaan konsep *accuracy* sehingga menghasilkan nilai yang merepresentasikan kualitas ejaan secara kuantitatif. Untuk rumusnya sebagai berikut:

$$skor\ ejaan = \frac{jumlah\ kata\ benar}{jumlah\ total\ kata} \times 100 \quad (4)$$

Sebagai contoh, apabila suatu teks terdiri dari 20 kata dan ditemukan 4 kesalahan ejaan, maka jumlah kata yang benar adalah 16 kata. Berdasarkan rumus skor ejaan dihitung sebagai berikut:

$$skor\ ejaan = \frac{16}{20} \times 100 = 80$$

Hasil perhitungan menunjukkan bahwa teks tersebut memiliki skor ejaan sebesar 80. Nilai skor yang lebih tinggi menunjukkan bahwa sebagian besar kata ditulis sesuai kaidah ejaan yang berlaku, sedangkan nilai yang lebih rendah menunjukkan adanya lebih banyak kesalahan penulisan dalam teks.

Pada penelitian ini, skor ejaan tidak hanya digunakan sebagai hasil evaluasi ketepatan ejaan, tetapi juga dimanfaatkan sebagai salah satu fitur dalam proses klasifikasi risiko literasi siswa. Dengan demikian, skor ejaan berfungsi sebagai representasi kuantitatif kemampuan menulis siswa yang selanjutnya dikombinasikan dengan nilai membaca dan nilai kuis untuk membentuk dataset klasifikasi.

E. Pembentukan Dataset dan Pelabelan

Setelah proses analisis kesalahan ejaan selesai dilakukan, tahap berikutnya adalah pembentukan dataset yang digunakan pada proses klasifikasi risiko literasi siswa. Dataset disusun dengan menggabungkan hasil analisis teks dan data akademik siswa. Fitur yang digunakan dalam penelitian ini terdiri atas skor ejaan, nilai membaca, dan nilai kuis.

Skor ejaan diperoleh dari hasil analisis kesalahan ejaan menggunakan metode Levenshtein Distance dan digunakan untuk merepresentasikan tingkat ketepatan penulisan siswa. Sementara itu, nilai membaca menggambarkan kemampuan siswa dalam memahami bacaan, sedangkan nilai kuis digunakan untuk merepresentasikan tingkat pemahaman siswa terhadap materi pembelajaran. Ketiga fitur tersebut dipilih karena mampu merepresentasikan aspek utama literasi, yaitu kemampuan menulis, membaca, dan pemahaman materi.

Setelah dataset terbentuk, setiap data diberikan label risiko literasi sebagai kelas target pada proses klasifikasi. Pelabelan dilakukan berdasarkan rata-rata nilai dari ketiga fitur yang digunakan. Nilai rata-rata yang diperoleh kemudian digunakan untuk menentukan kategori risiko literasi siswa, yaitu *low risk*, *moderate risk*, dan *high risk*. Kriteria pelabelan ditunjukkan pada Tabel 1 sebagai berikut.

Tabel 1. Pelabelan Dataset

Rata-rata nilai	Label
≥ 80	Low risk
60-79	Moderate risk
> 60	High risk

Kategori *low risk* menunjukkan siswa memiliki kemampuan literasi yang baik dengan risiko gangguan literasi yang rendah. Kategori *moderate risk* menunjukkan siswa memiliki kemampuan literasi yang cukup, namun masih memerlukan perhatian pada beberapa aspek pembelajaran. Adapun kategori *high risk* menunjukkan siswa memiliki kemampuan literasi yang relatif rendah sehingga memerlukan pendampingan dan intervensi lebih lanjut.

Proses pelabelan dilakukan sebelum tahap pelatihan model sehingga setiap data memiliki kelas target yang digunakan oleh algoritma Random Forest untuk mempelajari pola hubungan antara fitur masukan dan tingkat risiko literasi siswa.

F. Klasifikasi Menggunakan Random Forest

Random Forest merupakan salah satu algoritma *Machine Learning* berbasis *ensemble learning* yang bekerja dengan menggabungkan hasil prediksi dari banyak model untuk meningkatkan akurasi dan stabilitas klasifikasi [6]. Algoritma ini membangun sejumlah *decision tree* menggunakan subset data yang berbeda, kemudian mengombinasikan hasil prediksi setiap pohon untuk menghasilkan keputusan akhir [7]. Pendekatan tersebut memungkinkan Random Forest mengurangi risiko *overfitting* yang sering terjadi pada penggunaan *decision tree* tunggal serta meningkatkan kemampuan generalisasi model.

Menurut [8] proses kerja Random Forest pada penelitian ini meliputi tiga tahapan utama yaitu:

1. *Bootstrap sampling*, yaitu pembentukan beberapa subset data latih secara acak dengan metode *sampling with replacement*. Setiap subset yang terbentuk digunakan untuk melatih satu *decision tree* secara independen.
2. Pembentukan *decision tree*. Pada setiap simpul (*node*), algoritma memilih sebagian fitur secara acak untuk menentukan pemisahan data yang paling optimal. Pemilihan fitur secara acak bertujuan meningkatkan keberagaman antar pohon sehingga model tidak bergantung pada pola tertentu.
3. Agregasi prediksi (*ensemble aggregation*). Setiap *decision tree* menghasilkan prediksi terhadap data masukan, kemudian seluruh hasil prediksi digabungkan menggunakan metode *majority voting*. Kelas yang memperoleh suara terbanyak ditetapkan sebagai hasil klasifikasi akhir. Mekanisme ini membuat Random Forest lebih robust terhadap variasi data dan mampu menghasilkan prediksi yang lebih akurat.

G. Evaluasi Model

Evaluasi model merupakan tahapan penting dalam penelitian Machine Learning untuk menilai kemampuan model dalam melakukan klasifikasi secara akurat. Pada penelitian ini, kinerja model Random Forest dievaluasi menggunakan confusion matrix, yang merupakan metode evaluasi umum pada permasalahan klasifikasi.

Confusion matrix berfungsi untuk membandingkan hasil prediksi model dengan label sebenarnya pada data uji. Melalui matriks ini, dapat diketahui jumlah prediksi yang benar maupun salah pada setiap kelas, sehingga evaluasi tidak hanya berfokus pada ketepatan secara keseluruhan, tetapi juga pada kemampuan model dalam mengenali masing-masing tingkat risiko literasi dasar [9], [10]. Secara matematis, metrik evaluasi tersebut dihitung berdasarkan nilai pada confusion matrix dengan rumus sebagai berikut:

- 1) Accuracy = persentase prediksi yang benar.

$$Akurasi = \frac{TN + TP}{FP + TP + TN + FN} \quad (5)$$

- 2) Precision = tingkat ketepatan prediksi positif.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

- 3) Recall = kemampuan model mendeteksi data positif.

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

- 4) F1-Score = nilai rata-rata antara presisi dan recall.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

III. HASIL DAN PEMBAHASAN

Pengujian analisis kesalahan ejaan dilakukan untuk mengevaluasi kemampuan metode Levenshtein Distance dalam mengidentifikasi dan mengoreksi kata yang mengalami kesalahan penulisan. Proses pengujian dilakukan dengan membandingkan kata masukan dengan kata baku yang terdapat pada kamus acuan. Tingkat kemiripan kedua kata diukur menggunakan nilai *distance*, yang merepresentasikan jumlah operasi edit minimum yang diperlukan untuk mengubah kata masukan menjadi kata target. Semakin rendah nilai *distance* yang diperoleh, semakin tinggi tingkat kemiripan antara kedua kata tersebut.

Setelah proses deteksi dan koreksi kesalahan ejaan dilakukan menggunakan metode Levenshtein Distance, tahap selanjutnya adalah menghitung skor ejaan untuk setiap teks siswa. Skor ejaan digunakan untuk merepresentasikan tingkat ketepatan penulisan berdasarkan jumlah kata yang sesuai dengan kamus acuan dibandingkan dengan keseluruhan kata dalam teks. Nilai skor yang diperoleh berada pada rentang 0–100, di mana semakin tinggi skor menunjukkan semakin baik kualitas ejaan yang digunakan siswa. Hasil perhitungan skor ejaan pada beberapa data sampel ditunjukkan pada Tabel 2 berikut.

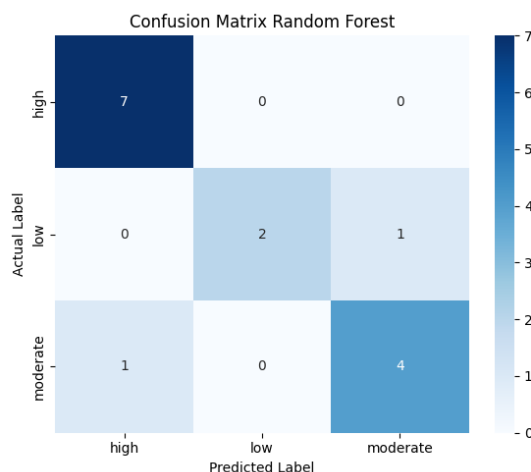
Tabel 2. Hasil perhitungan skor ejaan

No	Jumlah Kata	Kata Salah	Skor Ejaan
1	20	2	90
2	25	5	80
3	18	6	66,67

Hasil perhitungan pada Tabel 2, menunjukkan bahwa skor ejaan dapat digunakan untuk merepresentasikan tingkat ketepatan penulisan siswa secara kuantitatif. Nilai skor yang lebih tinggi menunjukkan jumlah kesalahan ejaan yang lebih sedikit, sedangkan peningkatan jumlah kesalahan penulisan akan menyebabkan penurunan skor ejaan secara proporsional.

Penggunaan skor ejaan memberikan ukuran yang lebih terstandarisasi dibandingkan hanya menghitung jumlah kesalahan ejaan. Dengan pendekatan ini, kualitas penulisan siswa dapat dievaluasi dan dibandingkan secara lebih objektif meskipun panjang teks atau jumlah kata yang ditulis berbeda.

Setelah proses pembentukan dataset dan pelabelan selesai dilakukan, tahap berikutnya adalah pelatihan model menggunakan algoritma Random Forest. Model dibangun menggunakan tiga fitur utama, yaitu skor ejaan hasil analisis Levenshtein Distance, nilai membaca, dan nilai kuis. Evaluasi kinerja model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengukur kemampuan model dalam mengklasifikasikan risiko literasi siswa ke dalam kategori *low risk*, *moderate risk*, dan *high risk*.



Gambar 1. Hasil Confusion Matrix

Berdasarkan Gambar 1 hasil confusion matrix, sebanyak 7 data kategori *high risk*, 2 data kategori *low risk*, dan 4 data kategori *moderate risk* berhasil diklasifikasikan dengan benar. Dengan demikian, jumlah prediksi yang benar adalah 13 data dari total 15 data pengujian. Nilai *accuracy* dihitung menggunakan persamaan (9):

$$accuracy = \frac{\text{Jumlah prediksi benar}}{\text{Jumlah data uji}} \quad (9)$$

Sehingga untuk substitusi nilai dihitung sebagai berikut:

$$accuracy = \frac{13}{15} \times 100 = 86,67\%$$

Nilai tersebut menunjukkan bahwa sebagian besar data berhasil diklasifikasikan sesuai dengan label yang dimiliki. Berdasarkan *classification report*, kategori *high risk* memperoleh nilai *precision* sebesar 0,88, *recall* sebesar 1,00, dan *F1-score* sebesar 0,93. Hasil ini menunjukkan bahwa seluruh siswa yang termasuk kategori risiko tinggi berhasil dikenali oleh model. Pada kategori *low risk*, model memperoleh nilai *precision* sebesar 1,00 dan *recall* sebesar 0,67, yang menunjukkan bahwa prediksi yang dihasilkan sangat tepat meskipun masih terdapat beberapa data yang belum berhasil teridentifikasi. Sementara itu, kategori *moderate risk* memperoleh nilai *precision*, *recall*, dan *F1-score* sebesar 0,80 yang menunjukkan performa klasifikasi yang cukup baik. Berdasarkan hasil *confusion matrix*, sebagian besar data berhasil diklasifikasikan dengan benar. Namun, masih terdapat beberapa kesalahan prediksi antara kategori *low risk* dan *moderate risk*, serta antara *moderate risk* dan *high risk*. Kondisi tersebut menunjukkan bahwa karakteristik

nilai antar kategori memiliki rentang yang relatif berdekatan sehingga batas klasifikasi tidak selalu dapat dipisahkan secara tegas. Fenomena ini umum terjadi pada kasus klasifikasi di bidang pendidikan karena kemampuan siswa berada pada spektrum yang bersifat kontinu.

Analisis feature importance dilakukan untuk mengetahui tingkat kontribusi masing-masing fitur terhadap proses klasifikasi. Pada algoritma Random Forest, nilai feature importance diperoleh dari akumulasi penurunan impurity (*Gini Impurity*) yang dihasilkan oleh setiap fitur pada seluruh pohon keputusan yang terbentuk. Semakin besar kontribusi suatu fitur dalam memisahkan data pada node-node pohon keputusan, semakin tinggi nilai importance yang diperoleh. Hasil perhitungan feature importance ditunjukkan pada Tabel 3 sebagai berikut

Tabel 3. Hasil Feature Importance

Fitur	Nilai Importance	Persentase
Nilai Baca	0,3627	36,27%
Nilai Kuis	0,3584	35,84%
Skor Ejaan	0,2789	27,89%

Berdasarkan Tabel 3, fitur nilai membaca memiliki nilai importance tertinggi sebesar 0,3627 atau 36,27%, diikuti oleh nilai kuis sebesar 0,3584 atau 35,84%, dan skor ejaan sebesar 0,2789 atau 27,89%. Hasil tersebut menunjukkan bahwa nilai membaca memberikan kontribusi paling besar dalam proses pengambilan keputusan model. Meskipun memiliki nilai importance yang lebih rendah, skor ejaan tetap memberikan pengaruh terhadap klasifikasi risiko literasi, sehingga dapat digunakan sebagai fitur pendukung dalam proses identifikasi risiko literasi siswa.

Temuan ini menunjukkan bahwa skor ejaan hasil analisis Levenshtein Distance tidak hanya berfungsi sebagai indikator kesalahan penulisan, tetapi juga dapat dimanfaatkan sebagai fitur pendukung dalam proses identifikasi risiko literasi. Dengan demikian, kemampuan menulis yang direpresentasikan melalui tingkat ketepatan ejaan memiliki keterkaitan dengan kemampuan literasi secara keseluruhan dan dapat digunakan sebagai salah satu indikator dalam deteksi dini risiko literasi siswa sekolah dasar.

Berdasarkan hasil penelitian, metode Levenshtein Distance memiliki kelebihan dalam mendeteksi kesalahan ejaan berbasis karakter, seperti pengurangan, penambahan, dan penggantian huruf. Metode ini juga mampu menghasilkan skor ejaan yang bersifat kuantitatif sehingga dapat dimanfaatkan sebagai fitur dalam klasifikasi risiko literasi. Selain itu, implementasinya relatif sederhana dan tidak memerlukan proses pelatihan model yang kompleks.

Namun, metode ini masih memiliki beberapa keterbatasan. Hasil analisis sangat bergantung pada kualitas dan kelengkapan kamus acuan yang digunakan. Selain itu, Levenshtein Distance hanya mengukur kemiripan karakter tanpa mempertimbangkan konteks kalimat, sehingga masih berpotensi menghasilkan koreksi yang kurang sesuai. Penentuan nilai *threshold* juga berpengaruh terhadap hasil analisis, di mana nilai yang terlalu kecil dapat menyebabkan kesalahan tidak terdeteksi, sedangkan nilai yang terlalu besar dapat menghasilkan kandidat koreksi yang kurang relevan. Di samping itu, kinerja model klasifikasi turut dipengaruhi oleh jumlah dan keragaman dataset yang digunakan selama proses pelatihan.

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, metode Levenshtein Distance berhasil diterapkan untuk menganalisis kesalahan ejaan pada teks tulisan siswa sekolah dasar. Metode ini mampu mendeteksi berbagai bentuk kesalahan penulisan, seperti pengurangan, penambahan, dan penggantian karakter, serta menghasilkan skor ejaan yang merepresentasikan tingkat ketepatan penulisan siswa secara kuantitatif. Skor ejaan yang dihasilkan kemudian dikombinasikan dengan nilai membaca dan nilai kuis sebagai fitur dalam klasifikasi risiko literasi menggunakan algoritma Random Forest. Hasil pengujian menunjukkan bahwa model mampu mengklasifikasikan tingkat risiko literasi siswa ke dalam kategori *low risk*, *moderate risk*, dan *high risk* dengan nilai *accuracy* sebesar 86,67%. Selain itu, model menunjukkan kemampuan yang baik dalam mengenali siswa pada kategori *high risk*, yang ditunjukkan oleh nilai *recall* sebesar 1,00.

Berdasarkan analisis *feature importance*, nilai membaca memberikan kontribusi terbesar dalam proses klasifikasi sebesar 36,27%, diikuti nilai kuis sebesar 35,84%, dan skor ejaan sebesar 27,89%. Meskipun memiliki kontribusi yang lebih rendah dibandingkan fitur lainnya, skor ejaan tetap memberikan pengaruh terhadap hasil klasifikasi dan dapat digunakan sebagai indikator pendukung dalam identifikasi risiko literasi siswa.

Secara keseluruhan, penelitian ini menunjukkan bahwa integrasi analisis kesalahan ejaan berbasis Levenshtein Distance dan klasifikasi menggunakan Random Forest dapat dimanfaatkan sebagai pendekatan untuk mendukung deteksi dini risiko literasi siswa sekolah dasar secara lebih objektif dan terukur.

UCAPAN TERIMA KASIH [JIKA ADA]

Penulis mengucapkan terima kasih kepada Instansi/perusahaan/lembaga yang telah memberikan dukungan financial terhadap penelitian ini.

DAFTAR PUSTAKA

- [1] I. Afkarina, A. V. Sutanto, and D. A. Wiliyanto, "PROFIL KEMAMPUAN LITERASI ANAK KELAS 3 SEKOLAH DASAR DI SURAKARTA," 2024.
- [2] D. G. S. Harahap, F. Nasution, E. S. Nst, and S. A. Sormin, "Analisis Kemampuan Literasi Siswa Sekolah Dasar," *Jurnal Basicedu*, vol. 6, no. 2, pp. 2089–2098, Feb. 2022, doi: 10.31004/BASICEDU.V6I2.2400.
- [3] A. Michael Bala Ledjap, F. Putri Rochmawati, D. Ayu Eka Marsanda, A. Puspita Sari, and U. Pembangunan Nasional Veteran Jawa Timur, "Pemanfaatan Natural Language Processing Untuk Pengecekan Ejaan Sesuai KBBI," *Jurnal Mahasiswa Teknik Informatika*, vol. 3, no. 2, pp. 46–56, Oct. 2024, doi: 10.35473/JAMASTIKA.V3I2.3255.
- [4] I. Oktavia Suzanti, Y. Dwi Putra Negara, A. Radiah, and A. Dwi Cahyani, "PENCARIAN ARTIKEL PARIWISATA BERBAHASA INDONESIA MENGGUNAKAN LEVENSHTTEIN DISTANCE QUERY SPELLING CORRECTION FOR INDONESIAN TOURISM ARTICLE USING LEVENSHTTEIN DISTANCE," vol. 13, no. 1, 2024.

- [5] I. F. Akbar, Mugiarto, K. F. Ramdhania, and Rasim, "Algoritma Levenshtein Distance sebagai Solusi Efisiensi Pencarian pada Training Registration System," *Journal of Students' Research in Computer Science*, vol. 5, no. 2, pp. 147–154, Nov. 2024, doi: 10.31599/mzr20n84.
- [6] L. Ratnawati and D. R. Sulistyaningrum, "Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel," 2019.
- [7] M. R. Inzaghi, E. R. Susanto, Amarudin, and Neneng, "Evaluasi Kinerja Model Random Forest Dalam Memprediksi Diabetes Berdasarkan Dataset Kesehatan di Indonesia," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 7, pp. 1857–1866, Jul. 2025, doi: 10.52436/1.jpti.871.
- [8] Y. Resti, C. Irsan, J. F. Latif, I. Yani, and N. R. Dewi, "A Bootstrap-Aggregating in Random Forest Model for Classification of Corn Plant Diseases and Pests," *Science and Technology Indonesia*, vol. 8, no. 2, pp. 288–297, Apr. 2023, doi: 10.26554/sti.2023.8.2.288-297.
- [9] A. Firizkiansah, A. Muhammad, and I. R. Maulana, "Optimasi Klasifikasi Data Teks Menggunakan Algoritma Logistic Regression dengan TF-IDF dan SMOTE," *JIKOMTI : Jurnal Ilmiah Ilmu Komputer dan Teknologi Informasi*, vol. 2, no. 1, pp. 29–36, May 2025, Accessed: Jan. 04, 2026. [Online]. Available: <https://ojs.sains.ac.id/index.php/Jikomti/article/view/97>
- [10] F. Reynaldi Valerian, M. Syarief, D. Abdul Fatah, J. Raya Telang, K. Kamal, and J. Timur, "KLASIFIKASI TINGKAT OBESITAS MENGGUNAKAN METODE GBM DAN CONFUSION MATRIX," 2025.