Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 5, No 1, November 2025
Hal 386−394
DOI 10.47065/comforch.v5i1.2345

Analisis Perbandingan Kemiripan Teks Bahasa Daerah Menggunakan Algoritma Naive Bayes dan K-Nearest Neighbor

Alfarizi[1,*], Herry Sujaini[2] , Niken Candraningrum[3]

Fakultas Teknik Informatika, Universitas Tanjungpura, Pontianak, Indonesia
Email: [1,*]alfaariizii2712@gmail.com, [2] hs@untan.ac.id , [3] nikenc@informatika.untan.ac.id
Email Penulis Korespondensi: alfaariizii2712@gmail.com

Abstrak−Indonesia, sebagai sebuah negara kepulauan, memiliki berbagai macam bahasa, Indonesia memiliki 718 bahasa daerah. Namun, banyak bahasa daerah yang menghadapi risiko penurunan pengguna hingga ternacam punah. Perkembangan teknologi membuka peluang untuk melakukan analisis pola dan karakteristik unik bahasa daerah melalui analisis n-gram yang menggunakan algoritma naive bayes dan k-nearest neighbor. Oleh karena itu, dilakukanlah penelitian ini dengan tujuan menganalisis kecenderungan kedekatan pola linguistik pada bahasa Jawa, Sunda, dan Melayu Pontianak menggunakan pendekatan n-gram sebagai salah satu upaya membantu pelestarian bahasa daerah di Indonesia. Penelitian ini tidak mengukur kemiripan bahasa secara langsung, melainkan memanfaatkan algoritma Naive Bayes dan k-Nearest Neighbor (k-NN) sebagai model klasifikasi, sehingga nilai kesalahan klasifikasi pada confusion matrix yang digunakan sebagai indikator melihat kemiripan distribusi fitur antarbahasa. Hasil menunjukkan bahwa Naive Bayes memberikan performa paling baik pada konfigurasi fitur Top 1% dengan nilai F1-score sebesar 0.922, melampaui k-NN yang memperoleh nilai maksimal 0.871. Analisis terhadap pola kesalahan menunjukkan bahwa pasangan bahasa dengan nilai kemiripan tertinggi adalah bahasa Jawa–Melayu (3.82%), yang diikuti bahasa Jawa–Sunda (2.33%), dan Melayu–Sunda (1.66%). Temuan ini mengindikasikan bahwa kemiripan yang muncul bersifat distribusional akibat kesamaan pola karakter, bukan kemiripan linguistik struktural. Dengan demikian, pendekatan klasifikasi dapat menjadi alat bantu untuk mengamati kecenderungan kedekatan pola teks antarbahasa daerah.

Kata Kunci: Bahasa Daerah; Naive Bayes; K-Nearest Neighbor

Abstract−*Indonesia, as an archipelagic country, has a wide variety of languages, with 718 regional languages. However, many regional languages are facing a decline in users and are threatened with extinction. Technological developments have opened up opportunities to analyze the patterns and unique characteristics of regional languages through n-gram analysis using naive bayes and k-nearest neighbor algorithms. Therefore, this study was conducted with the aim of analyzing linguistic pattern similarities in Javanese, Sundanese, and Pontianak Malay using the n-gram approach as an effort to help preserve regional languages in Indonesia. This study does not measure language similarity directly, but rather utilizes the Naive Bayes and k-Nearest Neighbor (k-NN) algorithms as classification models, so that the classification error value in the confusion matrix is used as an indicator to see the similarity of feature distribution between languages. The results show that Naive Bayes performs best on the Top 1% feature configuration with an F1-score of 0.922, surpassing k-NN which obtained a maximum score of 0.871. Analysis of error patterns shows that the language pairs with the highest similarity values are Javanese–Malay (3.82%), followed by Javanese–Sundanese (2.33%), and Malay–Sundanese (1.66%). These findings indicate that the similarities that appear are distributional due to similarities in character patterns, rather than structural linguistic similarities. Thus, the classification approach can be a tool for observing the tendency of text pattern proximity between regional languages.*

Keywords*: K-Nearest Neighbor, Naive Bayes, Regional Language*

1. PENDAHULUAN

Bahasa Indonesia, selain berfungsi sebagai identitas nasional, juga merupakan bahasa resmi negara yang terbentuk melalui proses perpaduan berbagai bahasa daerah yang ada di Indonesia. Keberagaman bahasa ini mencerminkan kekayaan budaya yang dimiliki oleh bangsa Indonesia. Berdasarkan data yang diperoleh dari Sensus Badan Pusat Statistik (BPS) tahun 2020, Indonesia tercatat memiliki sekitar 718 bahasa daerah yang tersebar di seluruh nusantara. Data dari sensus tersebut mencatat bahasa Jawa memiliki jumlah penutur terbanyak dengan sekitar 80 juta orang, diikuti oleh bahasa Sunda dengan 34 juta penutur. Keberagaman bahasa daerah ini menunjukkan betapa luasnya cakupan kebudayaan bahasa di Indonesia. Namun, di balik keanekaragaman tersebut, banyak bahasa daerah yang kini terancam punah, seiring dengan semakin berkurangnya jumlah penutur dan kurangnya perhatian terhadap upaya pelestarian bahasa tersebut.

Faktor modernisasi, migrasi, serta dominasi bahasa Indonesia sebagai bahasa komunikasi utama menyebabkan bahasa-bahasa daerah ini terpinggirkan dan semakin terlupakan. Salah satu pendekatan yang dapat diambil untuk melestarikan bahasa-bahasa daerah adalah dengan melakukan analisis kemiripan antarbahasa daerah. Analisis ini bertujuan untuk mengidentifikasi hubungan linguistik antara berbagai bahasa daerah, sehingga memudahkan dalam memahami perbedaan dan persamaan yang ada di dalamnya. Dalam hal ini, teknologi analisis teks menjadi sangat berguna untuk mengolah dan mengidentifikasi pola-pola linguistik yang ada dalam bahasa-bahasa tersebut.

Algoritma *Naive Bayes* dan *k-Nearest Neighbor* (k-NN) adalah dua metode yang sering digunakan dalam analisis klasifikasi data. Algoritma Naive Bayes bekerja dengan memanfaatkan probabilitas dan asumsi independensi antar fitur dalam teks untuk mengklasifikasikan bahasa berdasarkan ciri-cirinya. Sementara itu, algoritma *k- Nearest Neighbor* (k-NN) menggunakan pendekatan berbasis kedekatan antara data untuk mengelompokkan teks-teks yang memiliki kemiripan [1]. Kedua algoritma ini memiliki keunggulan dalam menganalisis pola-pola linguistik dan karakteristik bahasa daerah secara efisien. Pada penelitian lain yang telah dilakukan, menunjukkan bahwa algoritma Naive Bayes mampu memberikan akurasi hingga 88%, sedangkan algoritma *K-Nearest Neighbor* mampu mencapai akurasi sebesar 60% dalam klasifikasi teks bahasa daerah [2].

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 5, No 1, November 2025
Hal 386−394
DOI 10.47065/comforch.v5i1.2345

Bahasa Indonesia, selain berfungsi sebagai identitas nasional, juga merupakan bahasa resmi negara yang terbentuk melalui proses perpaduan berbagai bahasa daerah yang ada di Indonesia. Keberagaman bahasa ini mencerminkan kekayaan budaya yang dimiliki oleh bangsa Indonesia. Berdasarkan data Sensus Badan Pusat Statistik (BPS) tahun 2020, Indonesia tercatat memiliki sekitar 718 bahasa daerah yang tersebar di seluruh nusantara. Bahasa Jawa memiliki jumlah penutur terbanyak dengan sekitar 80 juta orang, diikuti oleh bahasa Sunda dengan 34 juta penutur. Keberagaman bahasa daerah ini menunjukkan betapa luasnya cakupan kebudayaan bahasa di Indonesia. Namun, di balik keanekaragaman tersebut, banyak bahasa daerah yang kini terancam punah seiring berkurangnya jumlah penutur dan minimnya upaya pelestarian bahasa daerah.

Faktor modernisasi, migrasi, serta dominasi bahasa Indonesia sebagai bahasa komunikasi utama menyebabkan bahasa-bahasa daerah semakin terpinggirkan. Salah satu cara untuk mendukung upaya pelestarian bahasa daerah adalah dengan melakukan analisis hubungan antarbahasa, termasuk melihat tingkat kemiripan linguistik yang terdapat di dalamnya. Analisis kemiripan ini dapat membantu memetakan keterhubungan antarbahasa serta mengidentifikasi karakteristik linguistik yang mungkin saling beririsan. Dalam konteks ini, teknologi analisis teks melalui pendekatan komputasional menjadi sangat relevan.

Algoritma Naive Bayes dan k-Nearest Neighbor (k-NN) adalah dua metode yang sering digunakan dalam analisis klasifikasi data. Algoritma Naive Bayes bekerja dengan memanfaatkan probabilitas dan asumsi independensi antar fitur dalam teks untuk mengklasifikasikan bahasa berdasarkan ciri-cirinya, sedangkan algoritma k-Nearest Neighbor (k-NN) menggunakan pendekatan berbasis kedekatan antara vektor data untuk mengelompokkan teks-teks yang memiliki kemiripan [1]. Kedua algoritma ini memiliki keunggulan dalam menganalisis pola linguistik dan karakteristik bahasa daerah secara efisien, terutama ketika representasi teks diubah ke dalam bentuk fitur n-gram. Representasi n-gram tersebut mengubah teks menjadi vektor frekuensi atau bobot karakter sehingga dapat diproses oleh kedua algoritma sebagai masukan numerik.

Meskipun demikian, Naive Bayes dan k-NN secara prinsip merupakan algoritma klasifikasi, bukan algoritma pengukuran kemiripan (similarity measure). Keduanya tidak secara langsung menghasilkan nilai kedekatan antarbahasa seperti Cosine Similarity atau Jaccard Similarity. Oleh karena itu, pada penelitian ini, kemiripan diukur melalui tingkat kesalahan klasifikasi pada confusion matrix, yaitu ketika suatu bahasa salah diprediksi sebagai bahasa lain. Frekuensi kesalahan tersebut merepresentasikan kedekatan pola linguistik antarbahasa berdasarkan pola n-gram yang diproses oleh model klasifikasi. Dengan demikian, analisis kemiripan yang dilakukan akan tetap relevan untuk melihat hubungan antarbahasa melalui perilaku model.

Pada penelitian lain yang telah dilakukan, menunjukkan bahwa algoritma Naive Bayes mampu memberikan akurasi hingga 88%, sedangkan algoritma k-Nearest Neighbor mampu mencapai akurasi sebesar 60% dalam klasifikasi teks bahasa daerah [2]. Meskipun performa ini telah banyak diteliti dalam konteks klasifikasi, penelitian terdahulu belum memanfaatkan pola misklasifikasi sebagai dasar untuk mengukur kemiripan antarbahasa. Dengan demikian, terdapat celah penelitian terkait pemanfaatan hasil confusion matrix untuk memetakan hubungan linguistik antarbahasa daerah yang dilakukan pada penelitian kali ini.

Pada penelitian yang dilakukan oleh Nurdin, Suhendri, Afrilia, dan Rizal (2021) membangun sistem yang dapat menentukan kategori tugas akhir mahasiswa yang ada di Program Studi Teknik Informatika Universitas Malikussaleh. Penelitian ini akan menentukan kategori tugas akhir berdasarkan pada bidang keahlian yaitu "Pengolahan Citra, Data Mining, Sistem Pengambilan Keputusan, Sistem Informasi Geografis, dan Sistem Pakar". Hasil dari penelitian menunjukkan 18 dari 20 data tugas akhir dapat diklasifikasikan dengan benar dengan rata-rata waktu proses pengujian selama 5,7406 detik per pengujian dengan nilai akurasi yang diklasifikasikan ke dalam 5 kelas sebesar 86.68% [3].

Penelitian yang dilakukan oleh R. D. Kurniawan & Muliawan (2024) mengatakan media sosial sangat berdampak besar dalam membangun sentimen dan preferensi politik publik mengambil media sosial sebagai studi kasus yaitu Twitter sehingga, *tweet* yang berkaitan dengan Pemilu dapat digunakan untuk melihat gambaran opini masyarakat. Tujuan dari penelitian yaitu mengklasifikasikan sentimen masyarakat di media sosial terkait Pemilu 2024 sebagai positif, netral, dan negatif. Data yang digunakan diambil pada Oktober hingga Desember 2023. Hasil akhir dari penelitian ini didapatkan sentimen positif sebanyak 78% yaitu 156 dari 200 data, sentimen netral sebesar 16% yang berjumlah 32 data, dan sentimen negatif dengan nilai 6% yang berjumlah 12 data [4].

Pada penelitian yang dilakukan oleh Winarti, Indriyawati, Vydia, dan Christanto menggunakan algoritma *naive bayes* karena algoritma ini dapat menghasilkan akurasi yang maksimal dengan data yang kecil. Sedangkan algoritma *k-nearest neighbor* digunakan karena algoritma ini kuat dari gangguan data. Hasil dari penelitian ini menunjukkan bahwa algoritma *naive bayes* memiliki kinerja yang lebih baik dengan akurasi sebesar 88% dibandingkan dengan *k-nearest neighbor* yang hanya mendapatkan akurasi sebesar 60% [2].

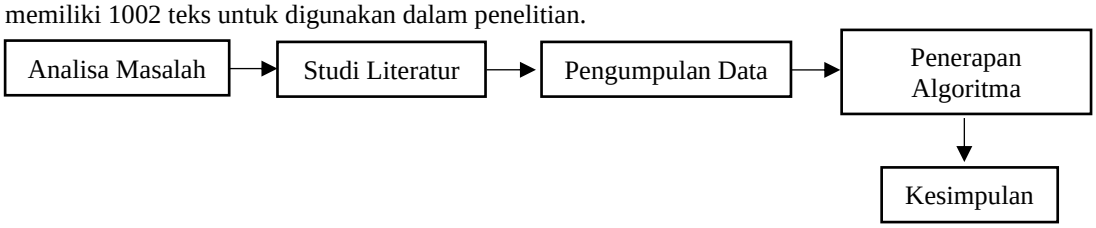
Penelitian ini bertujuan untuk menggali dan mengukur tingkat kemiripan bahasa-bahasa daerah di Indonesia, khususnya dalam konteks upaya pelestarian bahasa daerah. Dengan mengimplementasikan algoritma ini, kita dapat mengukur tingkat kemiripan antarbahasa daerah melalui analisis n-gram, sehingga dapat memperoleh gambaran yang lebih jelas tentang hubungan antar bahasa tersebut. Hasil dari analisis ini diharapkan dapat memberikan kontribusi dalam menjaga dan melestarikan bahasa-bahasa yang terancam punah. Dengan adanya pemahaman yang lebih mendalam tentang kemiripan dan perbedaan antarbahasa, diharapkan dapat muncul strategi yang lebih efektif dalam pelestarian bahasa daerah agar tetap lestari dan digunakan oleh generasi mendatang

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 5, No 1, November 2025
Hal 386−394
DOI 10.47065/comforch.v5i1.2345

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Metodologi penelitian ini diawali dengan mengidentifikasi permasalahan yang akan diteliti dan dilanjutkan dengan studi literatur untuk memahami teori - teori dari penelitian sebelumnya yang berkaitan dengan analisis kemiripan bahasa daerah dengan tujuan mendapatkan gambaran yang lebih luas mengenai topik penelitian[5],[6],[7],[8]. Penelitian menggunakan dua algoritma pembelajaran mesin, yaitu Naive Bayes dan k-Nearest Neighbor (k-NN), karena keduanya mampu mempelajari pola distribusi fitur n-gram pada teks. Dalam teori *pattern recognition*, kedekatan distribusi fitur antar kelas dapat menyebabkan model sering melakukan kesalahan klasifikaasi. Oleh karena itu, jika suatu bahasa sering diprediksi sebagai bahasa lain, maka hal ini dapat menjadi indikator adanya kemiripan pola linguistik antarbahasa. Pendekatan ini sejalan dengan penelitian terbaru yang menunjukkan bahwa struktur kesalahan (*misclassification pattern*) mengandung informasi mengenai hubungan antar kelas [9],[10],[11]. Setelah identifikasi masalah dan studi literatur dilakukan, dilanjutkan ke tahap pengumpulan data dengan mengumpulkan korpus teks bahasa daerah yang diperoleh dari NusaX. Korpus teks ini terdiri dari total 3006 teks, yang terbagi secara merata untuk masing-masing bahasa, dengan setiap bahasa memiliki 1002 teks untuk digunakan dalam penelitian.



Gambar 1. Tahapan Penelitian

2.2 *Naïve Bayes*

Naive Bayes adalah sebuah metode klasifikasi dalam ilmu data dan kecerdasan buatan yang bekerja dengan menilai seberapa besar kemungkinan suatu data termasuk ke dalam kategori tertentu berdasarkan pola dari data sebelumnya. Metode ini disebut "naive" karena mengasumsikan bahwa setiap fitur atau ciri dalam data bersifat independen satu sama lain, meskipun dalam dunia nyata asumsi tersebut tidak selalu benar [12],[13],[14],[15]. Meskipun demikian, pendekatan ini tetap sangat efektif dan sering memberikan hasil yang baik. Naive Bayes bekerja dengan cara menghitung seberapa sering sebuah kelas muncul, lalu melihat frekuensi kemunculan setiap ciri pada kelas tersebut. Berdasarkan pola tersebut, metode ini menentukan kelas mana yang paling mungkin untuk data baru [16],[17],[18]. Naive Bayes banyak digunakan dalam berbagai aplikasi, terutama pada klasifikasi teks seperti deteksi spam, analisis sentimen, dan pengelompokan dokumen karena prosesnya yang cepat, sederhana, serta mampu menangani jumlah data yang besar secara efisien.

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \propto P(C)\ P\ (X|C) \quad (1)$$

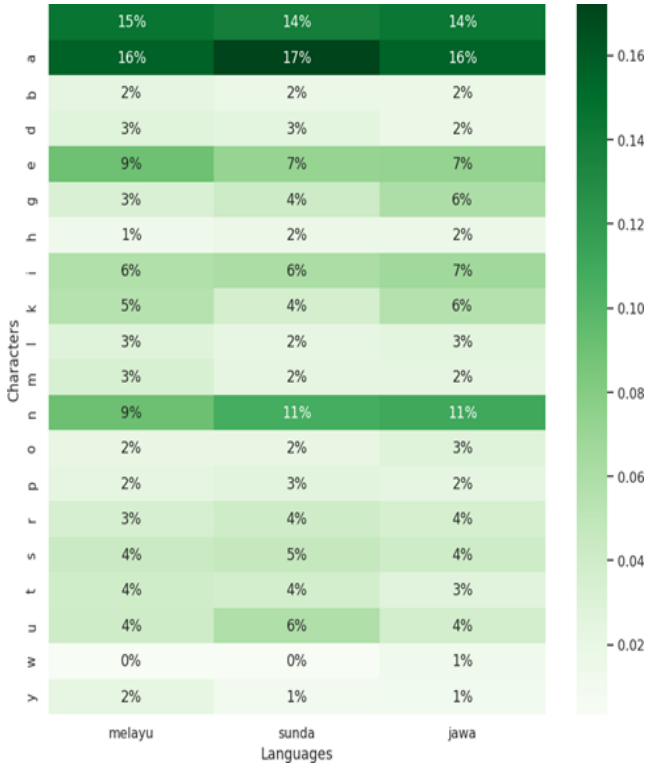
2.3 KNN *(K-Nearest Neighbors)*

K-Nearest Neighbors (KNN) adalah sebuah metode klasifikasi dan regresi yang bekerja berdasarkan kedekatan atau kemiripan data baru dengan data yang sudah ada. Ketika sebuah data baru ingin diklasifikasikan, KNN akan mencari sejumlah tetangga terdekat—biasanya disebut "K"—dari data tersebut dalam ruang fitur. Tetangga terdekat ini ditentukan berdasarkan jarak atau kemiripan, di mana data yang memiliki posisi paling dekat dianggap paling relevan. Setelah menemukan tetangga-tetangga terdekat tersebut, KNN menentukan kelas atau nilai data baru berdasarkan kelas atau nilai yang paling banyak muncul di antara tetangga tersebut. Karena metode ini tidak membangun model secara eksplisit, KNN termasuk teknik lazy learning, sehingga proses pelatihan sangat sederhana namun waktu prediksinya bisa lebih lama jika data besar. KNN sering digunakan dalam pengenalan pola, klasifikasi gambar, sistem rekomendasi, dan berbagai aplikasi lain yang membutuhkan penilaian berdasarkan kemiripan antar data [19],[20].

3. HASIL DAN PEMBAHASAN

Pada Gambar 2 berikut merupakan Penelitian yg menggunakan data sebanyak 3006 teks bahasa daerah yang terdiri dari tiga bahasa, yaitu bahasa Jawa Tengah, bahasa Sunda, dan bahasa Melayu Pontianak. Data tersebut diperoleh dari korpus NusaX, yang merupakan sebuah dataset paralel multibahasa yang mencakup berbagai macam bahasa daerah di Indonesia.

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 5, No 1, November 2025
Hal 386−394
DOI 10.47065/comforch.v5i1.2345



Gambar 2. Persebaran Fitur Bigram

Pada Gambar 1 adalah data persebaran fitur bigram. Setiap bahasa berisikan sebanyak 1002 data teks, sehingga total teks yang digunakan dalam penelitian ini adalah 3006. Dalam proses pengolahan data, didapatkan jumlah fitur unigram sebanyak 27 fitur. Kemudian untuk jumlah fitur bigram didapatkan sebanyak 432 fitur, dengan setiap bahasa memiliki fitur bigram sebanyak 403 fitur. Berikut Pada Tabel 1 merupakan fitur bigram yang didapatkan pada masing-masing bahasa.

Tabel 1. Persebaran Fitur Biagram

Bahasa	Fitur Bigram
Jawa	‘ a’, ‘ k’, ‘ n’, ‘ p’, ‘ s’, ‘a ’, ‘ak’, ‘an’, ‘ar’, ‘e ’, ‘en’, ‘g ’, ‘ga’, ‘i ’, ‘in’, ‘ka’, ‘ku’, ‘la’, ‘n ’, ‘na’, ‘ne’, ‘ng’, ‘ra’, ‘sa’, ‘si’
Sunda	‘ a’, ‘ d’, ‘ k’, ‘ n’, ‘ p, ‘ s’, ‘ t’, ‘a ’, ‘al’, ‘an’, ‘ar’, ‘di’, ‘en’, ‘eu’, ‘g ’, ‘ga’, ‘i ’, ‘ka’, ‘n ’, ‘na’, ‘ng’, ‘ra’, ‘sa’, ‘u ’, ‘un’
Melayu	‘ b’, ‘ d’, ‘ k’, ‘ m’, ‘ p’, ‘ s’, ‘ak’, ‘al’, ‘am’, ‘an’, ‘ar’, ‘as’, ‘at’, ‘da’, ‘e ’, ‘en’, ‘g ’, ‘i ’, ‘k ’, ‘ka’, ‘ma’, ‘n ’, ‘ng’, ‘ny’, ‘sa’, ‘se’, ‘u ’, ‘ya’

Pada tabel tersebut terdapat beberapa karakter bigram yang sama di antara ketiga bahasa seperti ‘ar’, ‘na’, ‘ka’. Hal tersebut menunjukkan adanya kemiripan di antara ketiga bahasa, yang tentunya dapat digunakan untuk analisis kemiripan ditahap selanjutnya.

Setelah fitur unigram dan bigram didapatkan dari masing – masing bahasa. Model *naïve bayes* dan *k-nearest* neighbor akan dilatih menggunakan data yang telah dibagi di tahap sebelumnya. Kemudian akan dilakukan pengujian model menggunakan data uji dan fitur-fitur yang telah ditentukan berdasarkan hasil uji coba fitur yang telah dilakukan. Berdasarkan hasil uji coba fitur, didapatkan bahwa fitur Top 1% dan Top 100 mendapatkan hasil yang maksimal dibandingkan dengan fitur lainnya, sehingga kedua fitur tersebut akan digunakan dalam proses analisis selanjutnya.

Pada fitur Top 100, algoritma *Naive Bayes* melakukan proses klasifikasi dengan memanfaatkan 100 fitur yang diperoleh dari masing-masing bahasa yang diteliti, yaitu bahasa Jawa, bahasa Sunda, dan bahasa Melayu Pontianak. Fitur ini diperoleh melalui analisis terhadap kata-kata atau token yang paling sering muncul dalam setiap bahasa. Setelah fitur ini diidentifikasi, fitur dari masing-masing bahasa kemudian digabungkan untuk mendapatkan kombinasi fitur unik yang mencerminkan karakteristik masing-masing bahasa.

Penggunaan fitur Top 100 ini bertujuan untuk menyaring elemen-elemen yang paling relevan dan berkontribusi pada proses klasifikasi, sehingga dapat meningkatkan akurasi dalam mengenali perbedaan antarbahasa. Selain itu, penggunaan fitur Top 100 ini juga berfungsi untuk meningkatkan efisiensi dalam proses klasifikasi data teks. Dengan mengurangi jumlah fitur yang digunakan hanya pada 100 fitur terpenting, algoritma menjadi lebih ringan dan lebih cepat dalam melakukan perhitungan, tanpa mengorbankan kualitas hasil klasifikasi.

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 5, No 1, November 2025
Hal 386−394
DOI 10.47065/comforch.v5i1.2345

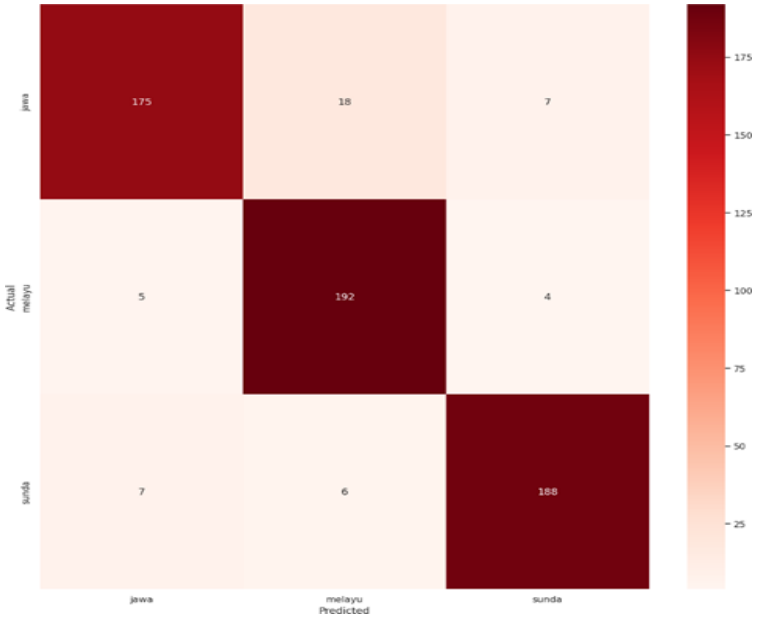
Untuk memperoleh nilai kemiripan antarbahasa, penelitian ini menggunakan data dari kesalahan klasifikasi yang muncul pada confusion matrix. Secara sederhana, kemiripan dihitung dengan melihat seberapa sering model salah mengenali teks dari satu bahasa sebagai bahasa lain, serta seberapa sering kesalahan yang sama terjadi pada arah sebaliknya. Kemudian kedua nilai kesalahan klasifikasi ini dijumlahkan dan dibandingkan dengan total data uji untuk mengetahui proporsi kesalahan yang merepresentasikan kedekatan pola linguistik antar dua bahasa. Dengan demikian, nilai kemiripan antarbahasa diperoleh menggunakan perhitungan. Berikut Hasil pengujian model dapat dilihat pada Tabel 2.

$$Kemiripan\ (A,B) = \frac{Teks\ bahasa\ A\ yang\ diprediksi\ sebagai\ bahasa\ B + Teks\ bahasa\ B\ yang\ diprediksi\ sebagai\ bahasa\ A}{Total\ data\ uji} \times 100\% \quad (1)$$

Tabel 2. Hasil Pengujian Model

No.	Algoritma dan Fitur	F1-Score
1	Naive Bayes Unigram	0.721
2	Naive Bayes Top 1%	0.922
3	Naive Bayes Top 100	0.915
4	kNN Unigram	0.668
5	kNN Top 1%	0.871
6	kNN Top 100	0.857

yang menunjukkan model *Naive Bayes* dengan Fitur Top 1% mendapatkan hasil yang paling tinggi dibandingkan dengan model dan fitur lain yang digunakan dalam pengujian. Kemudian untuk mengetahui nilai kemiripan antar bahasa, akan menggunakan hasil dari *confusion matrix*. Nilai kemiripan dihitung berdasarkan kesalahan prediksi yang dilakukan oleh model dalam melakukan klasifikasi. *Confusion matrix* dari model *naïve bayes* fitur top 1% dapat dilihat pada Gambar 3.



Gambar 3. Confusion Matrix Naive Bayes Fitur Top 1%

Pada Gambar 3 dapat dilihat beberapa kesalahan yang dilakukan oleh model *naïve bayes* pada fitur Top 1%. Kesalahan prediksi ini menunjukkan adanya kedekatan pola linguistik di antara ketiga bahasa, sehingga model tidak selalu mampu mengklasifikasikan teks sesuai dengan kelas aslinya. Berdasarkan gambar tersebut dapat diketahui bahwa:

a. Pada bahasa Jawa, 175 diprediksi dengan benar, sementara 18 diprediksi salah sebagai bahasa Melayu, dan 7 sebagai bahasa Sunda.
b. Pada bahasa Melayu, 192 diprediksi dengan benar, sementara 5 diprediksi salah sebagai bahasa Jawa, dan 4 sebagai bahasa Sunda.
c. Pada bahasa Sunda, 188 diprediksi dengan benar, sementara 7 diprediksi salah sebagai bahasa Jawa, dan 6 sebagai bahasa Melayu.

Berdasarkan hasil confusion matrix model Naïve Bayes dengan fitur Top 1%, nilai kemiripan antarbahasa diperoleh dengan menghitung jumlah kesalahan klasifikasi dua arah pada setiap pasangan bahasa. Kemiripan tertinggi terdapat pada pasangan Bahasa Jawa–Melayu, di mana teks Jawa yang salah diprediksi sebagai Melayu ditambah dengan teks Melayu yang salah diprediksi sebagai Jawa menghasilkan nilai kemiripan sebesar 3.82%. Perhitungan yang sama juga digunakan untuk pasangan bahasa lainnya, sehingga diperoleh nilai kemiripan Jawa–Sunda sebesar 2.33% dan

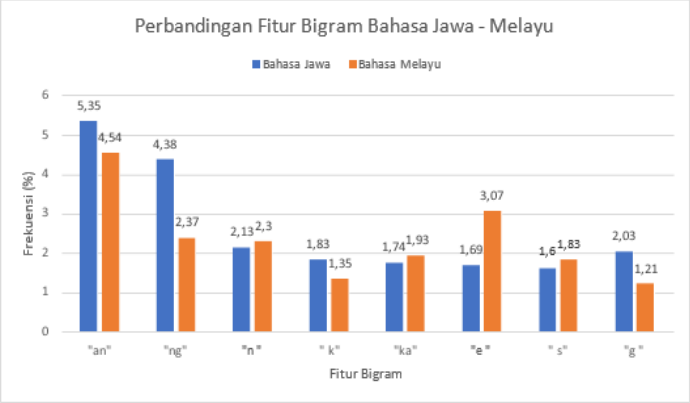
Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 5, No 1, November 2025
Hal 386−394
DOI 10.47065/comforch.v5i1.2345

Melayu–Sunda sebesar 1.66%. Berikut Pada Tabel 5, dapat dilihat 10 teratas kesalahan klasifikasi yang dilakukan oleh model naïve bayes pada fitur Top 1%.

Tabel 5. Top 10 Errors Naive Bayes Fitur Top 1%

No.	Expected	Predicted	Text
1.	Sunda	Melayu	Kfc Palembang Square Sering Janten Penyelamat Nalika Abdi Lapar Di Hotel
2.	Jawa	Melayu	Ternyata Kuwi Anake Temane
3.	Jawa	Melayu	Sundae Es Krime Nyegerake Lan Cukup Terjangkau Karo Papat Jinis Rasa
4.	Jawa	Melayu	Favorit Kita Yaiku Udang Mayonnaise
5.	Melayu	Sunda	Nyobe Pesan Shabu Shabu Menu Baru Di Gading Resto
6.	Jawa	Sunda	Telpon Ngalami Kesalahan Pas Sepisanan Diuripake
7.	Sunda	Melayu	Abdi Kaget Ku Palayanan Satpam Di Btpn Bromo Malang
8.	Melayu	Sunda	Layanan Pun Oke
9.	Jawa	Melayu	Menune Terjangkau Lan Cukup Bervariasi
10.	Jawa	Melayu	Kudu Ada Kontrol Kualitas Ing Departemen Rasa Supaya Kualitase Terjaga

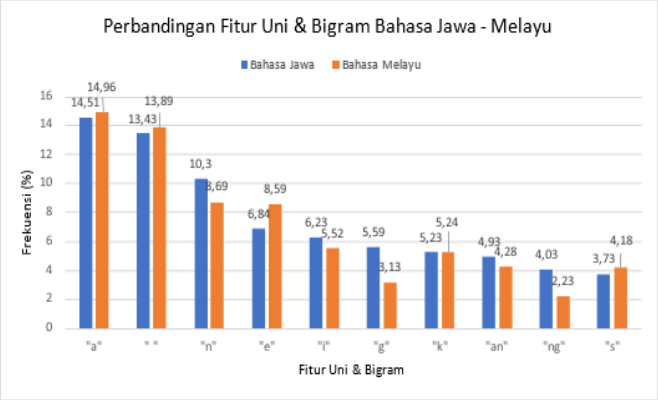
Analisis kemiripan antar bahasa juga dilihat berdasarkan kesalahan teratas klasifikasi teks yang dilakukan oleh model. Berdaarkan tabel tersebut, terdapat 5 teks bahasa Jawa yang diprediksi sebagai bahasa Melayu, 2 teks bahasa Sunda yang diprediksi sebagai bahasa Melayu, 2 teks bahasa Melayu yang diprediksi sebagai bahasa Sunda, dan 1 teks bahasa Jawa yang diprediksi sebagai bahasa Sunda. Pada kesalahan-kesalahan yang dilakukan oleh model dapat dilihat kesalahan prediksi paling banyak diprediksi sebagai bahasa Melayu, kesalahan tersebut dikarenakan pada beberapa teks menggunakan karakter ‘e’ sebagai akhiran dari suatu kata, yang berdasarkan frekuensi kemunculan karakter ‘e’ paling banyak muncul di bahasa melayu. Berikut Pada Gambar 4 dapat dilihat diagram frekuensi fitur bigram yang sering muncul di bahasa Jawa dan Melayu.



Gambar 4. Perbandingan Fitur Bigram Bahasa Jawa – Melayu

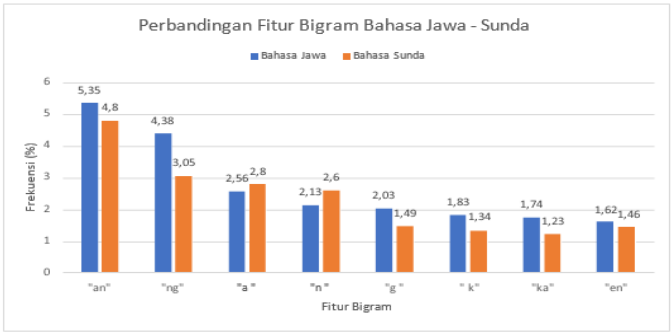
Persebaran fitur unigram dan bigram juga dapat digunakan untuk analisis kemiripan bahasa berdasarkan karakter yang sering digunakan oleh masing – masing bahasa. Pada gambar tersebut menunjukkan ada beberapa karakter yang frekuensi kemunculannya tidak berbeda jauh seperti karakter “n ”, “ka”, “ s”. Namun dari fitur yang sering digunakan pada bahasa Jawa, beberapa fitur memiliki frekuensi kemunculan rendah hingga tidak muncul di bahasa Melayu. Pada Gambar 5 juga dapat dilihat perbandingan frekuensi kemunculan untuk fitur unigram dan bigram pada bahasa Jawa dan Melayu.

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 5, No 1, November 2025
Hal 386−394
DOI 10.47065/comforch.v5i1.2345



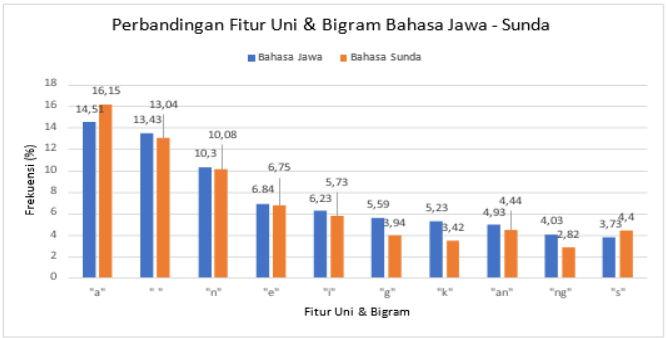
Gambar 5. Perbandingan fitur bigram bahasa Jawa – Melayu

Pada gambar tersebut menunjukkan terdapat beberapa karakter yang umum digunakan di bahasa Jawa juga digunakan di bahasa Melayu dengan frekuensi yang lebih tinggi, seperti karakter “a”, “e”, “k”, dan “s”. Frekuensi yang lebih tinggi tersebut menyebabkan model mengklasifikasikan beberapa teks bahasa Jawa ke bahasa Melayu. Pada Gambar 5 menunjukkan perbandingan fitur bigram antara bahasa Jawa dan Sunda.



Gambar 6. Perbandingan Fitur Bigram Bahasa Jawa – Sunda

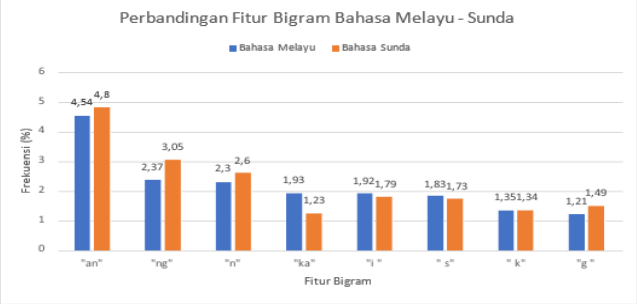
Pada gambar tersebut dapat dilihat frekuensi kemunculan bigram yang sama di antara kedua bahasa tidak terlalu mirip, hanya karakter “a ” dan “n ” yang frekuensi munculnya lebih banyak di bahasa Sunda dibandingkan di bahasa Jawa. Sementara karakter lain yang sering muncul di bahasa Jawa memiliki frekuensi kemunculan yang lebih banyak dibandingkan bahasa Sunda. Pada Gambar 7 dapat dilihat perbandingan fitur unigram dan bigram pada bahasa Jawa dan Sunda.



Gambar 7. Perbandingan Fitur Unigram dan Bigram Bahasa Jawa – Sunda

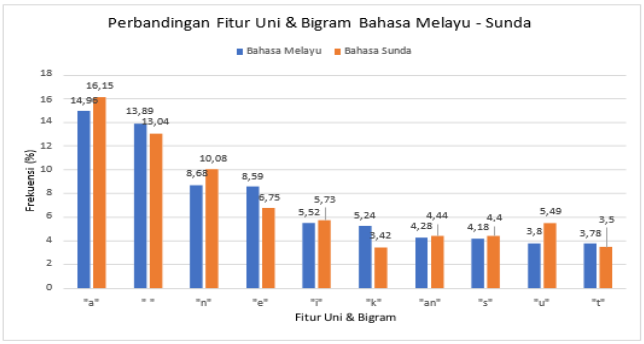
Pada gambar tersebut menunjukkan bahasa Sunda lebih banyak penggunaan karakter “a” dan “s” dibandingkan pada bahasa Jawa. Hal tersebut tentunya sedikit mempengaruhi model dalam melakukan klasifikasi teks bahasa Jawa yang banyak menggunakan karakter “a”, sehingga diklasifikasikan ke bahasa Sunda. Melalui pebandingan fitur yang digunakan oleh kedua bahasa juga dapat dilihat karakteristik dari kedua bahasa berdasarkan karakter yang sering digunakan. Pada Gambar 8 dapat dilihat perbandingan fitur bigram yang sering digunakan pada bahasa Melayu dan Sunda

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 5, No 1, November 2025
Hal 386−394
DOI 10.47065/comforch.v5i1.2345



Gambar 8. Perbandingan Fitur Bigram Bahasa Melayu – Sunda

dapat dilihat Pada gambar tersebut menunjukkan bahasa Sunda lebih banyak menggunakan karakter “an” dan “ng” dibandingkan bahasa Melayu. Sedangkan frekuensi pada karakter lain tidak berbeda jauh di antara kedua bahasa. Pada Gambar 9 dapat dilihat perbandingan frekuensi kemunculan fitur unigram dan bigram yang digunakan oleh bahasa Melayu dan Sunda



Gambar 9. Perbandingan Fitur Unigram dan Bigram Bahasa Melayu – Sunda

Gambar tersebut menunjukkan penggunaan karakter “a”, “n”, dan “u”pada bahasa Melayu tidak sebanyak yang digunakan pada bahasa Sunda. Frekuensi tersebut menunjukkan karakteristik bahasa Sunda lebih sering menggunakan karakter seperti “a”, “n”, dan “u” dalam membentuk teks bahasa Sunda. Berdasarkan hasil perbandingan fitur unigram dan bigram pada ketiga bahasa, terlihat bahwa kemiripan yang muncul terutama dipengaruhi oleh adanya tumpang tindih pola n-gram karakter yang digunakan pada teks masing-masing bahasa. Bahasa Jawa dan Melayu menunjukkan banyak kesamaan pada bigram seperti “an”, “ra”, “ka”, dan “na”, sehingga distribusi n-gram keduanya menjadi lebih mirip dan menyebabkan model lebih sering melakukan kesalahan prediksi pada pasangan bahasa tersebut. Sebaliknya, Bahasa Jawa–Sunda maupun Melayu–Sunda memiliki lebih sedikit n-gram yang muncul dengan frekuensi tinggi secara bersamaan, sehingga model lebih jarang keliru dalam membedakan kedua pasangan bahasa tersebut. Dengan demikian, kemiripan antarbahasa yang dihasilkan oleh struktur n-gram yang saling tumpang tindih. Hal ini menjadi faktor utama yang membuat model menganggap beberapa bahasa lebih dekat satu sama lain, yang kemudian tercermin dalam pola kesalahan klasifikasi pada confusion matrix.

4. KESIMPULAN

Melalui penelitian yang telah dilakukan, didapatkan hasil bahwa algoritma Naive Bayes terbukti lebih efektif dan efisien untuk digunakan dalam mengklasifikasikan teks bahasa daerah dibandingkan dengan algoritma k- Nearest Neighbor (k-NN). Dalam penelitian ini, analisis menggunakan confusion matrix menunjukkan adanya kemiripan antarbahasa yang diteliti, yaitu bahasa Jawa, Sunda, dan Melayu Pontianak, meskipun tingkat kemiripan antarbahasa tersebut bervariasi. Hasil penelitian juga menunjukkan bahwa algoritma Naive Bayes pada fitur Top 1% mendapatkan nilai F1- score tertinggi sebesar 0.922, yang menunjukkan kinerja yang sangat baik dalam melakukan klasifikasi bahasa daerah. Sementara itu, algoritma k-NN meskipun memberikan hasil yang cukup baik, memiliki nilai F1- score yang lebih rendah dibandingkan dengan Naive Bayes, dengan nilai maksimal sebesar 0,871. Berdasarkan kinerja algoritma yang paling baik, didapatkan nilai kemiripan tertinggi di antara ketiga bahasa adalah bahasa ‘Jawa – Melayu’ sebesar 3.82%, kemudian antara bahasa ‘Jawa – Sunda’ sebesar 2.33%, dan nilai kemiripan terkecil antara bahasa ‘Melayu – Sunda’ dengan nilai sebesar 1.66%. Nilai kemiripan tersebut dapat terlihat dari adanya kesamaan pada fitur unigram dan bigrams yang digunakan oleh masing – masing bahasa, seperti karakter ‘e’ yang sering muncul di bahasa Jawa, namun lebih dominan muncul di bahasa Melayu. Kemudian karakter ‘n’ yang sering muncul di bahasa Jawa dan Melayu, tapi lebih dominan muncul di bahasa Sunda. Nilai kemiripan antar bahasa ini tentunya dipengaruhi oleh banyak faktor seperti jumlah data dan pengolahan data yang dilakukan. Berdasarkan hasil penelitian kali ini, membuktikan bahwa nilai kemiripan antara bahasa Jawa, Sunda, dan Melayu terbilang cukup rendah atau tidak begitu mirip antar ketiga bahasa. Serta dalam analisis kemiripan bahasa daerah,

algoritma *naive bayes* lebih baik untuk digunakan dibandingkan dengan algoritma *k- nearest neighbor*. Selain itu, berdasarkan temuan penelitian dan mengacu pada potensi pengembangan yang masih terbuka, terdapat beberapa saran yang dapat diterapkan untuk penelitian selanjutnya. Pertama, penggunaan dataset yang lebih besar dapat meningkatkan representativitas dan akurasi hasil klasifikasi. Kedua, peningkatan kualitas pra-pemrosesan data diperlukan agar fitur yang digunakan lebih bersih, relevan, dan konsisten, sehingga dapat mengurangi potensi bias dalam proses pelatihan model. Ketiga, penelitian lanjutan dapat mempertimbangkan penggunaan metrik evaluasi tambahan atau algoritma lain yang lebih kompleks untuk memberikan perspektif yang lebih komprehensif terkait kedekatan atau perbedaan pola linguistik antarbahasa daerah.

REFERENCES

[1] M. Yudhi Putra and D. Ismiyana Putri, "Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI," *J. Tekno Kompak*, vol. 16, no. 2, pp. 176–187, 2022, doi: 10.33365/jtk.v16i2.2002

[2] T. Winarti, H. Indriyawati, V. Vydia, and F. W. Christanto, "Performance comparison between naive bayes and k-nearest neighbor algorithm for the classification of indonesian language articles," *IAES Int. J. Artif. Intell.*, vol. 10, no. 2, pp. 452–457, 2021, doi: 10.11591/IJAI.V10.I2.PP452-457.

[3] N. Nurdin, M. Suhendri, Y. Afrilia, and R. Rizal, "Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (NBC)," *Sist. J. Sist. Inf.*, vol. 10, no. 2, pp. 268–279, 2021, doi: 10.32520/stmsi.v10i2.1193.

[4] R. D. Kurniawan and J. Muliawan, "Sentiment Analysis of Indonesian Election 2024 Using the K-Nearest Neighbor Method," *J. Tek. Inform.*, vol. 5, no. 3, pp. 653–659, 2024, doi: 10.52436/1.jutif.2024.5.2.1934

[5] D. C. Agustin, M. A. Rosid, and N. Ariyanti, "Implementasi Convolutional Neural Network Untuk Deteksi Kesegaran Pada Apel," *J. Fasilkom*, vol. 13, no. 02, pp. 145–150, 2023, doi: 10.37859/jf.v13i02.5175.

[6] S. Afolabi, N. Ajadi, A. Jimoh, and I. Adenekan, "Predicting Diabetes Using Supervised Machine Learning Algorithms," *Res. Sq.*, Jun. 2024, doi: 10.21203/rs.3.rs-4527374/v1.

[7] M. Amin, "Bahasa Melayu Dalam Tradisi Islam Nusantara," *J. Islam. Soc. Sci.*, vol. 2, no. 2, pp. 64–77, 2021, doi : 10.30821/islamijah.v2i1.17080

[8] L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-2941-7.

[9] A. Irawan, J. Ahyar, and M. Mahsa, "Pemertahanan Bahasa Jawa di Tengah Masyarakat Multilingual Kecamatan Cot Girek," *J. Yudistira Publ. Ris. Ilmu Pendidik. dan Bhs.*, vol. 2, no. 4, pp. 368–385, 2024, doi: https://doi.org/10.61132/yudistira.v2i4.1202.

[10] D. Kurniawan, *Pengenalan Machine Learning dengan Python*, 1st ed. Jakarta: Elex Media Komputindo, 2020.

[11] Y. I. Kurniawan, "Perbandingan Algoritma Naive Bayes dan C.45 dalam Klasifikasi Data Mining," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 455–464, Oct. 2018, doi: 10.25126/jtiik.201854803.

[12] N. Lailiyah and F. Indri Wijayanti, "Kekerabatan Bahasa Jawa, Bali, dan Bima: Perspektif Linguistik Historis Komparatif," *Linguist. Indones.*, vol. 40, no. 2, pp. 327–345, 2022.

[13] O. Mailani, I. Nuraeni, S. A. Syakila, and J. Lazuardi, "Bahasa Sebagai Alat Komunikasi Dalam Kehidupan Manusia," Online, 2022.

[14] A. Purwanto and E. A. Darmadi, "Perbandingan Minat Siswa Smu Pada Metode Klasifikasi Menggunakan 5 Algoritma," *J. Komput. Dan Inform.*, vol. 2, no. 1, pp. 43–47, 2018.

[15] H. Sujaini and A. Bijaksana Putra, "Analysis of language identification algorithms for regional Indonesian languages," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, p. 1741, 2024, doi: 10.11591/ijai.v13.i2.pp1741-1752.

[16] S. S. Utama, A. W. Nuswantoro, A. Febrianto, and S. Mulyono, "Hubungan Kekerabatan Bahasa Jawa dan Bahasa Melayu (Kajian Linguistik Historis Komparatif)," *J. Pendidikan, Bhs. dan Budaya*, vol. 2, no. 3, pp. 60–76, 2023.

[17] G. P. Papadopoulos and G. E. Chryssolouris, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies*, vol. 12, no. 7, p. 113, 2024.

[18] Anugrah, I. G. (2021). Penerapan Metode N-Gram dan Cosine Similarity Dalam Pencarian Pada Repositori Artikel Jurnal Publikasi. Building of Informatics, Technology and Science (BITS), 3(3), 275–284. https://doi.org/10.47065/bits.v3i3.1058

[19] Nugraha, S. D., Putri, R. R. M., & Wihandika, R. C. (2017). Penerapan Fuzzy K- Nearest Neighbor (FK-NN) Dalam Menentukan Status Gizi Balita. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 1(9), 925–932.

[20] Nurhidayat, R., & Dewi, K. E. (2023). Penerapan Algoritma K-Nearest Neighbor Dan Fitur Ekstraksi N-Gram Dalam Analisis Sentimen Berbasis Aspek. Jurnal Ilmiah Komputer Dan Informatika, 12(1), 91–100. https://doi.org/10.34010/komputa.v12i1.9458