ANALISIS PERBANDINGAN KEMIRIPAN TEKS BAHASA DAERAH DI INDONESIA MENGGUNAKAN ALGORITMA NAIVE BAYES DAN K-NEAREST NEIGHBOR

SKRIPSI

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:
ALFARIZI
NIM D1041191018

FAKULTAS TEKNIK
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Alfarizi

NIM : D1041191018.

menyatakan bahwa dalam skripsi yang berjudul "ANALISIS PERBANDINGAN KEMIRIPAN TEKS BAHASA DAERAH DI INDONESIA MENGGUNAKAN ALGORITMA NAIVE BAYES DAN K-NEAREST NEIGHBOR" tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi manapun. Sepanjang pengetahuan Saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademis dan hukum di kemudian hari apabila pernyataan yang dibuat ini tidak benar.

Pontianak, 30 Januari 2025

Alfarizi
NIM D1041191018

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS TANJUNGPURA
FAKULTAS TEKNIK
Jalan Prof. Dr. H. Hadari Nawawi Pontianak 78124
Telepon (0561) 740186, WA: +62821522809907
Email : ft@untan.ac.id Website : http://teknik.untan.ac.id

HALAMAN PENGESAHAN

ANALISIS PERBANDINGAN KEMIRIPAN TEKS BAHASA DAERAH DI INDONESIA MENGGUNAKAN ALGORITMA NAIVE BAYES DAN K-NEAREST NEIGHBOR

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:

Alfarizi
NIM D1041191018

Telah dipertahankan di depan Penguji Skripsi pada tanggal 30 Januari 2025
dan diterima sebagai salah satu persyaratan untuk memperoleh gelar sarjana.

Susunan Penguji Skripsi:

Ketua,

Prof. Dr. Herry Sujaini, S.T., M.T.
NIP 196806291997021001

Penguji Utama,

Rina Septiriana, S.T., M.Cs.
NIP 198709232020122001

Sekretaris,

Niken Candraningrum, S.T., M.Cs.
NIP 199309272022032010

Penguji Pendamping,

Dr. Arif Bijaksana Putra Negara, S.T., M.T.
NIP 197208081998021002

Pontianak, 30 Januari 2025
Dekan,

Dr.-Ing. Ir. Slamet Widodo, M.T., IPM
NIP 196712231992031002

HALAMAN PERSEMBAHAN

Segala puji dan syukur kepada Allah SWT atas berkat dan rahmat-Nya skripsi ini dapat terselesaikan. Shalawat dan salam juga tercurah kepada junjungan dan suri tauladan kita Nabi Muhammad Shalallahu Alaihi Wassalam.

Tidak ada halaman yang paling penting dalam tugas akhir ini melainkan halaman persembahan. Tugas akhir ini tentunya saya persembahkan kepada orang-orang tercinta dan tersayang. Terima kasih sebesar-besarnya saya persembahkan kepada:

1. Kedua orang tua tercinta, Bapak Drs. Alfian, MM. dan Ibu Nursila, yang selalu memberikan doa, motivasi serta dukungan, baik secara moril maupun materil selama pengerjaan skripsi ini.
2. Kakak dan abang, Muftia dan Alghifari yang senantiasa memotivasi dan menyemangati setiap langkah yang dijalankan dalam proses penyelesaian skripsi.
3. Teman-teman terdekat penulis "Eraldo, Gabe, Yudha, Fadil, Sheva", yang setia menemani penulis dalam proses penulisan tugas akhir ini, serta selalu memberi semangat untuk segera menyelesaikan tahap terakhir dalam perkuliahan.
4. Teman-teman "The Chosen 9 Naga", yang selalu memberikan semangat, motivasi, serta canda tawa dalam setiap pertemuan. Tanpa hal tersebut tentu semuanya terasa tidak mudah untuk dilalui.
5. Teman-teman Informatika Angkatan 2019, yang selalu mendukung dan senantiasa penulis jadikan motivasi untuk dapat menyelesaikan skripsi ini.
6. Teman online penulis, yang selalu memberi pelajaran kepada penulis untuk menjadi pribadi yang lebih baik setiap harinya, serta selalu menyemangati penulis sepanjang penulisan skripsi.
7. Seluruh *coffeeshop* yang telah penulis jadikan tempat menulis sepanjang proses penulisan skripsi.
8. *Last but not least, I wanna thank me*.

KATA PENGANTAR

Puji syukur kehadirat Allah Subhanahu wa ta'ala atas kehendak serta pertolongan-Nya maka penelitian dan penulisan skripsi yang berjudul "Analisis Perbandingan Kemiripan Teks Bahasa Daerah di Indonesia Menggunakan Algoritma Naive Bayes dan K-Nearest Neighbor" dapat terselesaikan dengan baik. Penelitian ini mendapatkan bantuan dari berbagai pihak yang terlibat. Oleh karena itu, penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada Bapak Rudy Dwi Nyoto, S.T., M.Eng. selaku Dosen Pembimbing Akademik, Bapak Professor Herry Sujaini, S.T., M.T. selaku Dosen Pembimbing I yang senantiasa memberikan bimbingan dan arahan selama proses penulisan skripsi, Ibu Niken Candraningrum, S.T., M.Cs. selaku Dosen Pembimbing II yang telah memberikan bimbingan, arahan, dan masukan dalam penulisan skripsi, Ibu Rina Septiriana, S.T., M.Cs. selaku Dosen Penguji I yang telah memberi saran dan masukan untuk penulisan skripsi, dan Bapak Dr. Arif Bijaksana Putra Negara, S.T., M.T. selaku Dosen Penguji II yang telah memberikan masukan dalam pengerjaan skripsi. Penulis berharap penelitian ini dapat bermanfaat bagi berbagai pihak yang membutuhkan. Penulis juga menyadari bahwa setiap karya penelitian tentu tidak terlepas dari kekurangan dan kesalahan, sehingga penulis mengharapkan adanya masukan, kritik, serta saran yang membangun demi perbaikan yang lebih baik untuk penelitian ini dan penelitian di masa yang akan datang.

Pontianak, 30 Januari 2025
Penulis,

Alfarizi

ABSTRAK

Sebagai sebuah negara kepulauan, Indonesia 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 kemiripan bahasa daerah, terutama bahasa Jawa Tengah, Sunda, dan Melayu Pontianak sebagai salah satu upaya membantu pelestarian bahasa daerah di Indonesia. Hasil analisis kemiripan antar bahasa dihitung berdasarkan kesalahan pada *confusion matrix* dan kinerja dari algoritma akan dinilai menggunakan metrik akurasi dan F1-*score*. Algoritma *naive bayes* dengan fitur gabungan unigram dan bigrams menunjukkan kinerja yang paling baik dengan nilai akurasi dan F1-*score* sebesar 0.921. Hasil penelitian menunjukkan nilai kemiripan tertinggi pada bahasa 'Jawa – Melayu' meskipun hanya sebesar 3.82% dan terendah pada bahasa 'Melayu – Sunda' sebesar 1.66%. Nilai kemiripan tersebut didasarkan pada karakter yang dominan muncul di suatu bahasa seperti 'e' pada bahasa Melayu serta 'a' dan 'u' pada bahasa Sunda. Penelitian ini membuktikan bahwa hanya sedikit kemiripan yang ada di antara bahasa Jawa, Sunda, dan Melayu.

Kata kunci: bahasa daerah, *naive bayes*, *k-nearest neighbor*, *confusion matrix*. F1-*score*.

ABSTRACT

As an archipelagic country, Indonesia has a variety of languages, Indonesia has 718 regional languages. However, many regional languages are at risk of decreasing users to the point of extinction. Technological developments open up opportunities to analyze patterns and unique characteristics of regional languages through n-gram analysis using the naive bayes and k-nearest neighbor algorithms. Therefore, this study was conducted with the aim of analyzing the similarities of regional languages, especially Central Javanese, Sundanese, and Pontianak Malay as an effort to help preserve regional languages in Indonesia. The results of the analysis of similarities between languages are calculated based on errors in the confusion matrix and the performance of the algorithm will be assessed using accuracy and F1-score metrics. The naive bayes algorithm with combined unigram and bigrams features showed the best performance with an accuracy and F1-score value of 0.921. The results of the study showed the highest similarity value in the 'Javanese - Malay' language although only 3.82% and the lowest in the 'Malay - Sundanese' language of 1.66%. The similarity value is based on the dominant characters that appear in a language such as 'e' in Malay and 'a' and 'u' in Sundanese. This study proves that there are only a few similarities between Javanese, Sundanese, and Malay.

Keywords: regional language, naive bayes, k-nearest neighbor, confusion matrix. F1-score.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR KODE PROGRAM

BAB I
PENDAHULUAN

1.1 Latar Belakang

Bahasa Indonesia, selain menjadi identitas nasional, juga merupakan bahasa resmi Negara Indonesia yang terbentuk melalui perpaduan berbagai bahasa daerah yang ada di Indonesia. Sebagai negara kepulauan, Indonesia sungguh kaya akan keragaman, mulai dari suku, ras, budaya, agama, hingga bahasa. Dengan keberagaman ini, hampir setiap daerah di Indonesia memiliki bahasa daerahnya sendiri dengan ciri khas, ragam, dan logat yang berbeda. Proses migrasi, perdagangan, serta interaksi antarbudaya yang terjadi di masa lalu tentunya berpengaruh terhadap bahasa daerah yang ada di Indonesia. Hal ini menyebabkan adanya kemiripan di antara bahasa daerah dalam berbagai aspek, seperti kosakata, karakteristik, tata bahasa, dan penulisan.

Seiring perkembangan zaman, banyak masyarakat yang dalam percakapan sehari-harinya mencampurkan bahasa daerah dengan bahasa Indonesia, hal tersebut membuat bahasa tertentu semakin dikenal oleh banyak orang. Namun, di sisi lain, beberapa bahasa daerah di Indonesia mulai mengalami risiko penurunan penggunaan bahkan terancam punah. Untuk menghadapi situasi tersebut, mengetahui kemiripan bahasa daerah dapat digunakan untuk meningkatkan pemahaman antarbudaya di Indonesia. Pendalaman terkait kemiripan antar bahasa juga merupakan salah satu upaya yang dapat dilakukan dalam melestarikan keragaman bahasa daerah di Indonesia.

Berdasarkan *long form* sensus yang dilakukan oleh Badan Pusat Statistik pada tahun 2020, Indonesia memiliki 718 bahasa daerah dengan berbagai keunikan dari berbagai aspeknya (Aziz, 2023). Dari banyaknya bahasa daerah di Indonesia, tentunya tidak dapat dimungkiri bahwa ada kemiripan dari bahasa-bahasa tersebut yang dipengaruhi beberapa hal seperti letak geografis, serapan kata, dan penggunaan sehari-hari antar

masyarakat. Berdasarkan data l*ong form* sensus tahun 2020, bahasa daerah dengan jumlah penutur terbanyak adalah bahasa Jawa dengan jumlah penutur sebanyak 80 juta penutur. Pada urutan kedua, ditempati oleh bahasa Sunda dengan jumlah penutur sebanyak 34 juta penutur. Sementara itu, untuk bahasa Melayu memiliki jumlah penutur sebanyak 19 juta penutur. Pada penelitian ini terdapat 3 bahasa daerah yang akan diteliti kemiripannya yaitu bahasa Jawa, bahasa Sunda, dan bahasa Melayu.

Seiring dengan kemajuan teknologi, metode komputasional telah menjadi alat yang efisien untuk menganalisis pola dan karakteristik bahasa. Penelitian kali ini akan menggunakan algoritma *naive bayes* dan *k-nearest neighbor* (k-NN) dalam melakukan analisis kemiripan bahasa melalui analisis n-gram. Analisis n-gram digunakan untuk melihat pola karakter yang sering digunakan oleh setiap bahasa. Algoritma *naive bayes* dan *k-nearest neighbor* cukup mudah untuk diimplementasikan dan memiliki performa yang baik pada sumber daya terbatas maupun sumber daya besar. Dalam penelitian yang dilakukan oleh (Yudhi Putra & Ismiyana Putri, 2022), kedua algoritma tersebut cukup umum digunakan dalam klasifikasi data, dimana algoritma *k-nearest neighbor* melakukan proses menemukan pola baru dalam data dengan menghubungkan jarak antara data yang sudah ada dengan data baru, sedangkan algoritma *naive bayes* menggunakan probabilitas dalam melakukan klasifikasi, algoritma ini menggunakan data yang sudah ada dalam menentukan kelas untuk data baru. Algoritma ini juga sering digunakan untuk studi kasus *data mining* maupun *text mining*. Pada penelitian yang dilakukan oleh (Sujaini & Bijaksana Putra, 2024), didapatkan hasil bahwa algoritma *naive bayes* memiliki kinerja paling baik dan disusul oleh algoritma *k-nearest neighbor* dalam melakukan tugas identifikasi 8 bahasa daerah di Indonesia. Pada penelitian lain yang dilakukan oleh (Winarti et al., 2021), kedua algoritma memiliki nilai akurasi yang cukup tinggi dengan nilai akurasi sebesar 88% untuk algoritma *naive bayes* dan 60% untuk algoritma *k-nearest neighbor*.

Berdasarkan pemaparan yang telah dijelaskan di atas, diharapkan dengan dilakukannya penelitian kali ini dapat memberikan pemahaman yang lebih mendalam terkait hubungan antar bahasa melalui angka kemiripan bahasa daerah, mengetahui pola dan karakteristik unik bahasa daerah melalui analisis n-gram, serta membantu pelestarian bahasa daerah yang ada di Indonesia untuk generasi yang akan datang.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang sudah dijelaskan, dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana kinerja algoritma *naive bayes* dan *k-nearest neighbor* dalam menghitung nilai kemiripan teks bahasa daerah di Indonesia?
2. Bagaimana nilai kemiripan bahasa daerah di Indonesia menggunakan algoritma *naive bayes* dan *k-nearest neighbor*?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah mendapatkan hasil kinerja algoritma *naive* bayes dan *k-nearest neighbor* dalam menghitung nilai kemiripan teks bahasa daerah serta mendapatkan hasil nilai kemiripan bahasa daerah di Indonesia.

1.4 Pembatasan Masalah

Agar proses pengerjaan penelitian ini lebih terarah, maka dibuat batasan masalah pada penelitian ini, diantaranya yaitu:

1. Bahasa yang akan digunakan dalam penelitian ini adalah bahasa Jawa Tengah, Sunda, dan Melayu Pontianak.
2. Hasil dari penelitian ini yaitu nilai kemiripan teks bahasa daerah yang disajikan dalam bentuk *confusion matrix.*

1.5 Sistematika Penulisan

Laporan tugas akhir ini akan disajikan dalam lima bab dengan sistematika pembahasan sebagai berikut:

BAB I : LATAR BELAKANG

Bab ini berisi tentang latar belakang masalah, perumusan masalah,

tujuan penelitian, pembatasan masalah, serta sistematika penulisan pembuatan tugas akhir ini.

BAB II : TINJAUAN PUSTAKA

Bab ini berisi landasan teori terkait penelitian yang akan dilakukan sebagai penunjang dalam pengerjaan tugas akhir.

BAB III : METODOLOGI PENELITIAN

Pada bab ini membahas tentang langkah-langkah yang akan dilakukan dalam penelitian, mulai dari identifikasi masalah, studi literatur, pengumpulan data, pra-pemrosesan data, pelatihan model, pengujian model, evaluasi model, analisis hasil, dan pengambilan kesimpulan.

BAB IV : HASIL DAN PEMBAHASAN

Pada bab ini berisi penjelasan mengenai nilai kemiripan dari bahasa daerah yang menggunakan algoritma naive bayes dan k-nearest neighbor.

BAB V : PENUTUP

Pada bab ini berisi kesimpulan penelitian mengenai nilai kemiripan dan kinerja algoritma yang lebih baik digunakan dalam analisis nilai kemiripan bahasa daerah di Indonesia, serta saran yang dapat digunakan untuk penelitian-penelitian selanjutnya.