ANALISIS PERBANDINGAN KEMIRIPAN TEKS BAHASA DAERAH DI INDONESIA MENGGUNAKAN ALGORITMA *RANDOM FOREST* DAN *SUPPORT VECTOR MACHINE*

SKRIPSI

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:
FATHIMAH NUR UMMAMI
NIM D1041201063

FAKULTAS TEKNIK
UNIVERSITAS TANJUNGPURA
PONTIANAK
2024

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Fathimah Nur Ummami

NIM : D1041201063

menyatakan bahwa dalam skripsi yang berjudul "Analisis Perbandingan Kemiripan Teks Bahasa Daerah di Indonesia Menggunakan Algoritma *Random Forest* dan *Support Vector Machine*" 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, 5 Desember 2024

Fathimah Nur Ummami
NIM D1041201063

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN,
RISET, DAN TEKNOLOGI
UNIVERSITAS TANJUNGPURA
FAKULTAS TEKNIK
Jalan Prof. Dr. H. Hadari Nawawi Pontianak 78124 Telp. (0561) 740186
Kotak Pos 1049 Email: ft@untan.ac.id Website: http:/Teknik.untan.ac.id

HALAMAN PENGESAHAN

ANALISIS PERBANDINGAN KEMIRIPAN TEKS BAHASA DAERAH DI INDONESIA MENGGUNAKAN ALGORITMA *RANDOM FOREST* DAN *SUPPORT VECTOR MACHINE*

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:

Fathimah Nur Ummami
NIM D1041201063

Telah dipertahankan di depan Penguji Skripsi pada tanggal 5 Desember 2024 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,

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

Sekretaris,

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

Penguji Pendamping,

Helen Sastypratiwi, S.T., M.Eng.
NIP 198601172012122004

Pontianak, 5 Desember 2024
Dekan,

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

HALAMAN PERSEMBAHAN

Terima kasih untuk diri sendiri yang telah memulai apa yang sudah dipilih dan menyelesaikannya dengan baik. Tentu tidak mudah untuk sampai dititik ini, mungkin langkah ini tidak sepenuhnya mendapatkan beberapa peran dikehidupan layaknya seperti anak yang lain, namun langkah ini tetap terus berjalan dan kuat karena diiringi dengan doa-doa tulus dari mama. Skripsi ini saya persembahkan kepada orang-orang yang sangat saya sayangi dan cintai.

Teruntuk mama dan ayah, dengan segala hormat cinta, kasih sayang, dan terima kasih saya ucapkan yang sebesar-besarnya. Saya persembahkan skripsi ini sebagai tanda tanggung jawab dalam menjalani apa yang sudah menjadi pilihan menuju jalan masa depan. Terima kasih selalu memberikan doa yang tak pernah putus, pengorbanan, dan terima kasih atas pelajaran hidup yang mungkin bisa jadi bekal untuk kedepannya.

Terima kasih penulis ucapkan kepada sahabat-sahabat yang senantiasa selalu menemani dan memberikan support tiada habisnya yaitu Sherina Tri Manggiri, Fiyen Febryan, Evelyn Oktaviana Hutapea, Salsabil Humaira, Tari Wahyuni, Cik Dara Kamila, Jesica Fitriani Silalahi, Putri Sintang Anugeraeni, dan Tri Rizky Saputri.

Teruntuk pembimbing akademik selama masa perkuliahan yaitu ibu Helen Sastypratiwi, S.T., M.Eng., dengan segala hormat dan terima kasih saya ucapkan kepada ibu yang sudah memberikan bimbingan, arahan, dan ilmu yang berharga kepada saya selama masa perkuliahan.

Teruntuk pembimbing tugas akhir, dengan segala hormat dan terima kasih sebesar-besarnya saya ucapkan kepada Bapak Prof. Dr. Herry Sujaini, S.T., M.T, dan ibu Rina Septiriana, S.T., M.Cs. Terima kasih kepada bapak dan ibu yang telah memberikan segala bantuan, bimbingan, nasihat, ilmu, dan kesabaran yang telah bapak dan ibu berikan selama menjadi pembimbing saya yang akan selalu saya ingat. Saya mohon maaf apabila ada kesalahan yang saya perbuat selama masa bimbingan.

Terima kasih juga kepada teman seperjuangan di masa menempuh perkuliahan yaitu Fika Tsalsabila Tinanda dan Rendy Kurniawan, serta terima kasih kepada teman-teman seperjuangan INFORMATIKA 20 dan teman-teman SETARA TEKNIK 20.

KATA PENGANTAR

Puji dan Syukur penulis panjatkan atas kehadirat Allah SWT, yang telah memberikan Rahmat serta karunia-Nya, Sholawat serta salam yang senantiasa dilimpahkan kepada junjungan Nabi Muhammad SAW, sehingga penulis dapat menyelesaikan skripsi ini dengan judul "Analisis Perbandingan Kemiripan Teks Bahasa Daerah di Indonesia Menggunakan Algoritma *Random Forest* dan *Support Vector Machine*". Sebagai syarat untuk menyelesaikan Program Sarjana (S1) pada Fakultas Teknik, Jurusan Informatika, Universitas Tanjungpura.

Penulis menyadari bahwa dalam penulisan skripsi ini tidak terlepas dari bantuan berbagai pihak, baik dari segi moril dan materil. Oleh karena itu, penulis menyampaikan rasa hormat dan terima kasih yang sebesar-besarnya kepada pihak pihak yang telah turut andil memberikan bantuan dan dukungan kepada penulis. Secara khusus penulis menyampaikan ucapan terima kasih kepada Bapak Prof. Dr. Herry Sujaini, S.T., M.T. selaku dosen Pembimbing I dan Ibu Rina Septiriana, S.T., M.Cs. selaku dosen Pembimbing II, serta kepada Bapak Dr. Arif Bijaksana Putra Negara, S.T., M.T. selaku dosen Penguji I dan Ibu Helen Sastypratiwi, S.T., M.Eng. selaku dosen Penguji II.

Penulis menyadari bahwa skripsi ini belum sempurna dan masih terdapat kekurangan. Oleh karena itu, dengan segenap kerendahan hati penulis mengharapkan saran dan kritik yang membangun dari semua pihak untuk perbaikan dan pengembangan penulisan berikutnya. Semoga hasil penelitian ini dapat memberikan kontribusi yang bermanfaat bagi pembelajaran dan pemahaman mendalam, terutama mengenai Analisis Perbandingan Kemiripan Teks Bahasa Daerah di Indonesia Menggunakan Algoritma *Random Forest* dan *Support Vector Machine*.

Pontianak, 5 Desember 2024
Penulis,

Fathimah Nur Ummami

ABSTRAK

Indonesia memiliki keberagaman bahasa daerah, termasuk di Kalimantan Barat yang terdiri dari beberapa bahasa daerah seperti bahasa Dayak Iban, Dayak Pesaguan, Melayu Pontianak, dan Melayu Sambas. Kemiripan bahasa dalam satu rumpun sering kali menyebabkan kesalahan klasifikasi dalam sistem pengenalan bahasa otomatis. Oleh karena itu, penelitian ini dilakukan bertujuan untuk membandingkan performa algoritma *Random Forest* dan *Support Vector Machine* (SVM) dalam mengklasifikasikan teks bahasa daerah serta mengidentifikasi tingkat kemiripan antar bahasa yang digunakan dalam penelitian. Metodologi yang diterapkan pada penelitian meliputi pengumpulan data dari korpus Nusantara, bahasa yang digunakan dalam penelitian ini yaitu bahasa Dayak Iban (997 data), bahasa Dayak Pesaguan (1603 data), bahasa Indonesia (9099 data), bahasa Melayu Pontianak (1747 data), dan bahasa Melayu Sambas (9099 data). Selanjutnya dilakukan pra-pemrosesan teks (*cleaning*, *case folding*, tokenisasi, dan vektorisasi TF-IDF), serta penanganan ketidakseimbangan data menggunakan teknik SMOTE. Model kemudian dikembangkan menggunakan dua pendekatan, yaitu data Non-SMOTE dan data SMOTE, dengan algoritma *Random Forest* dan SVM. Evaluasi model dilakukan dengan menggunakan *heatmap confusion matrix* untuk menganalisis kesalahan klasifikasi serta menghitung *accuracy* model. Hasil penelitian menunjukkan bahwa algoritma SVM (SMOTE) mencapai *accuracy* tertinggi sebesar 95,36%, sementara *Random Forest* (SMOTE) memperoleh *accuracy* sebesar 94,18%. Dari analisis kesalahan klasifikasi, ditemukan bahwa pasangan bahasa Indonesia dan Melayu Sambas memiliki tingkat kemiripan tertinggi sebesar 6,45%. Kesimpulannya, algoritma SVM (SMOTE) lebih baik dibandingkan *Random Forest* dalam mengklasifikasikan bahasa daerah di Kalimantan Barat.

Kata Kunci: Kemiripan Bahasa, *Random Forest*, *Support Vector Machine*, SMOTE, Bahasa Daerah

ABSTRACT

Indonesia has a diverse range of regional languages, including those in West Kalimantan, which consists of several local languages such as Dayak Iban, Dayak Pesaguan, Malay Pontianak, and Malay Sambas. The similarity of languages within the same linguistic family often leads to misclassification in automatic language recognition systems. Therefore, this study aims to compare the performance of the Random Forest and Support Vector Machine (SVM) algorithms in classifying regional language texts and identifying the degree of similarity among the languages used in the research. The methodology applied in this study includes data collection from the Nusantara corpus. The languages used in this study are Dayak Iban (997 data samples), Dayak Pesaguan (1,603 data samples), Indonesian (9,099 data samples), Malay Pontianak (1,747 data samples), and Malay Sambas (9,099 data samples). The next steps involve text preprocessing (cleaning, case folding, tokenization, and TF-IDF vectorization) and handling data imbalance using the SMOTE technique. The model was then developed using two approaches: Non-SMOTE data and SMOTE data, with Random Forest and SVM algorithms. Model evaluation was conducted using a heatmap confusion matrix to analyze classification errors and calculate model accuracy. The results show that the SVM (SMOTE) algorithm achieved the highest accuracy of 95.36%, while Random Forest (SMOTE) obtained an accuracy of 94.18%. From the classification error analysis, it was found that the Indonesian and Malay Sambas language pair had the highest similarity rate of 6.45%. In conclusion, the SVM (SMOTE) algorithm outperforms Random Forest in classifying regional languages in West Kalimantan.

***Keywords**: Language Similarity, Random Forest, Support Vector Machine, SMOTE, Regional Languages*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

KODE PROGRAM

BAB I
PENDAHULUAN

1.1 Latar Belakang

Di Indonesia, terdapat 718 bahasa daerah yang digunakan oleh penduduk dari berbagai suku dan etnis yang tersebar di seluruh kepulauan Nusantara. Meskipun lebih dari 90% penduduk Indonesia memahami dan menggunakan Bahasa Indonesia sebagai bahasa komunikasi sehari-hari, bahasa ini bukanlah bahasa ibu bagi sebagian besar masyarakat. Sebaliknya, mayoritas penduduk menggunakan salah satu dari ratusan bahasa daerah tersebut sebagai bahasa ibu mereka. Penggunaan bahasa sehari-hari di Indonesia sering kali bersifat campuran, di mana Bahasa Indonesia yang formal bercampur dengan berbagai dialek Melayu atau bahasa daerah lainnya. Hal tersebut mengakibatkan sulitnya dalam mengenali kemiripan bahasa satu dengan bahasa lainnya (Purnamalia et al., 2023).

Provinsi Kalimantan Barat, wilayah ini memiliki keberagaman budaya dan bahasa, dengan kehadiran berbagai suku seperti Dayak Iban, Dayak Pesaguan, Melayu Pontianak, Melayu Sambas, dan lainnya. Setiap suku memiliki bahasa daerahnya masing-masing yang khas dan mencerminkan identitas budaya mereka. Namun, di tengah keragaman ini terdapat kemiripan antara bahasa-bahasa tersebut, terutama yang berasal dari rumpun bahasa yang sama, seperti Melayu. Namun dari rumpun Dayak juga terdapat kemiripan antara bahasa-bahasa tersebut. Untuk rumpun Melayu adanya kemiripan bahasa antara Melayu Pontianak dengan Melayu Sambas, sedangkan rumpun dayak adanya kemiripan bahasa antara Dayak Iban dengan Dayak Pesaguan. Kemiripan ini sering kali menyebabkan tumpang tindih linguistik, di mana bahasa-bahasa yang digunakan sehari-hari tampak saling berbaur, membuat identifikasi bahasa tertentu menjadi lebih rumit. Kondisi ini semakin diperburuk oleh kebiasaan masyarakat yang secara alami mencampuradukkan elemen dari berbagai bahasa dalam percakapan sehari-hari, terutama di daerah dengan kontak budaya yang intens antar suku dan etnis.

Dalam konteks ini, klasifikasi bahasa menjadi langkah penting untuk mengelompokkan suatu bahasa secara efektif, sehingga dapat diterapkan untuk

mengenali suatu bahasa tertentu dengan lebih akurat. Proses ini melibatkan pengumpulan teks kalimat dari suatu bahasa daerah tertentu, yang kemudian akan dianalisis dan dikelompokkan berdasarkan informasi label yang diberikan. Sehingga suatu teks yang sudah dilabel dapat dikelompokkan berdasarkan label-label tersebut. Menurut (Wibawa et al., 2018) terdapat beberapa metode klasifikasi bahasa, di antaranya adalah pendekatan pembelajaran mesin seperti *Random Forest*, *Support Vector Machine* (SVM), *Naive Bayes Claasifier* (NBC) dan *K-Nearest Neighbors* (KNN). *Random Forest* merupakan perkembangan dari metode CART, yang menggabungkan teknik *bootstrap aggregating* (*bagging*) dan seleksi fitur acak. Konsep ini melibatkan pembentukan banyak pohon secara bersamaan, membentuk suatu hutan, dan kemudian melakukan analisis terhadap kumpulan pohon tersebut (Syukron & Subekti, 2018). *Support Vector Machine* (SVM) merupakan sebuah metode yang digunakan untuk melakukan prediksi, baik dalam konteks klasifikasi maupun regresi (Darmayanti et al., 2018). *Naïve Bayes Classifier* (NBC) merupakan salah satu metode yang digunakan untuk mengklasifikasikan sekumpulan teks. Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya (Setiawan et al., 2015). Algoritma K-*Nearest Neighbor* adalah metode yang termasuk dalam kategori pembelajaran terawasi (*supervised learning*). Dalam pembelajaran terawasi, fokusnya adalah untuk menemukan pola baru dalam data dengan mengaitkan pola yang sudah ada dengan data yang baru (Susandi et al., 2016).

Data yang dikumpulkan dalam penelitian ini tidak seimbang, maka dari itu data akan diseimbangkan menggunakan teknik SMOTE. *Synthetic Minority Oversampling Technique* (SMOTE) merupakan algoritma yang bekerja dengan tujuan mengontrol pemerataan distribusi data pada suatu kelas dalam dataset dengan membuat sampel buatan dari kelas minoritas (Sharfina Nabilah & Ramadhan Nur, 2022). *Confusion Matrix* merupakan teknik yang digunakan untuk mengevaluasi klasifikasi model untuk memperkirakan objek yang benar atau salah. Sebuah matriks dari prediksi akan dibandingkan dengan kelas asli yang berisi informasi aktual dan prediksi nilai klasifikasi (Pravina et al., 2019).

Penelitian mengenai klasifikasi bahasa sudah pernah dilakukan oleh (Tuhenay & Mailoa, 2021), peneliti melakukan perbandingan klasifikasi bahasa menggunakan metode NBC dan SVM dengan hasil yaitu kedua metode tersebut bagus dalam melakukan identifikasi bahasa, namun SVM lebih efektif daripada NBC. Kemudian penelitian yang dilakukan oleh (Yasir & Suraji, 2023) yaitu peneliti membandingkan akurasi beberapa metode klasifikasi, seperti *Naïve Bayes*, *Decision Tree*, dan *Random Forest*. Hasil dari penelitian tersebut menunjukkan bahwa akurasi *Naïve Bayes* mencapai 90%, *Decision Tree* 83%, dan *Random Forest* 87%. Berdasarkan penelitian-penelitian tersebut, pada penelitian ini peneliti memilih algoritma *Random Forest* dan *Support Vector Machine* (SVM) dalam menganalisis kemiripan bahasa. Kedua algoritma tersebut digunakan untuk mengidentifikasi kemiripan antara bahasa-bahasa yang berbeda dan melakukan analisis perbandingan mengenai kemiripan bahasa.

Berdasarkan masalah yang ada dalam penelitin ini, maka peneliti menggunakan beberapa bahasa berdasarkan suku yang ada di Kalimantan Barat. Beberapa bahasa yang memiliki kemiripan bahasa tersebut seperti dari rumpun dayak yaitu Bahasa Dayak Iban, Bahasa Dayak Pesaguan. Bahasa dari rumpun melayu yaitu Bahasa Melayu Pontianak, dan Bahasa Melayu Sambas. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi penting dalam memahami keragaman bahasa di Indonesia, terutama dalam konteks mengenali kemiripan bahasa daerah di Kalimantan Barat.

1.2 Perumusan Masalah

Berdasarkan pada latar belakang yang telah dijelaskan, berikut beberapa rumusan masalah yang dapat menjadi fokus dalam penelitian ini:

1. Bagaimana hasil dari perbandingan antara algoritma *Random Forest* dan *Support Vector Machine*?
2. Bagaimana tingkat kemiripan antara bahasa Indonesia, bahasa Dayak Iban, bahasa Dayak Pesaguan, bahasa Melayu Pontianak, dan bahasa Melayu Sambas?

1.3 Tujuan Penelitian

Penelitian ini bertujuan mengetahui hasil perbandingan algoritma *Random Forest* dan SVM serta mengetahui tingkat kemiripan antara bahasa Indonesia, bahasa Dayak Iban, bahasa Dayak Pesaguan, bahasa Melayu Pontianak, dan bahasa Melayu Sambas.

1.4 Pembatasan Masalah

Penulis mengusulkan pembatasan masalah penelitian sebagai berikut:

1. Data yang digunakan berupa teks dari bahasa Indonesia dan beberapa bahasa daerah yaitu bahasa Melayu Pontianak, bahasa Melayu Sambas, bahasa Dayak Iban, dan bahasa Dayak Pesaguan.
2. Data yang digunakan yaitu data Non-SMOTE dan data SMOTE.
3. Algoritma yang digunakan yaitu *Random Forest* dan *Support Vector Machine* (SVM).
4. Hasil dan analisis dari penelitian ini didapatkan dari hasil *heatmap confusion matrix*.

1.5 Sistematika Penulisan

Sistematika dari penulisan skripsi ini terdiri dari lima bab yaitu, Bab I Pendahuluan, Bab II Tinjauan Pustaka, Bab III Metode Penelitian, Bab IV Hasil dan Pembahasan, dan Bab V Penutup.

BAB I PENDAHULUAN

Berisi tentang latar belakang, rumusan masalah, batasan masalah, tujuan

penelitian, metode penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI

Berisi tentang teori-teori mengenai *Natural Language Processing (NLP)*, *Term Frequency-Inverse Document Frequency* (TF-IDF), Bahasa Daerah, Imbalance Class, *Synthetic Minority Over-sampling Technique* (SMOTE), algoritma *Random Forest*, algoritma *Support Vector Machine, dan Confusion Matrix*.

BAB III METODE PENELITIAN

Berisi tentang diagram alir dan rangkaian alur tahapan pelaksanaan penelitian, dari mulai pemahaman masalah yang diangkat oleh penulis, bagaimana persiapan *dataset*, sampai pada pembuatan model.

BAB IV HASIL DAN PEMBAHASAN

Berisi tentang hasil skenario pengujian yang sudah dilakukan. Dari hasil tersebut akan dibahas mengenai kinerja dari algoritma serta hasil kemiripan berdasarkan perhitungan algoritma *Random Forest* dan *Support Vector Machine*, beserta arah kesimpulan penelitian.

BAB V PENUTUP

Berisi tentang kesimpulan serta saran terkait penelitian yang sudah dilakukan oleh penulis.

BAB II
TIN JAUAN PUSTAKA

2.1 Penelitian Terkait

Dalam penelitian ini terdapat hal yang mengacu kepada beberapa penelitian yang telah dilakukan oleh peneliti-peneliti sebelumnya dengan topik terkait. Hal-hal yang mencangkup adalah klasifikasi bahasa, algoritma *Random Forest*, algoritma *Support Vector Machine* (SVM), dan *Cunfusion Matrix*.

(Tuhenay & Mailoa, 2021) melakukan penelitian perbandingan klasifikasi bahasa menggunakan metode *Naive Bayes Classfier* (NBC) dan *Support Vector Machine* (SVM). Peneliti bertujuan untuk mengidentifikasi bahasa dalam bentuk teks. Melalui penulisan ini akan dipaparkan hasil identifikasi bahasa Indonesia, Ambon dan Jawa dengan menggunakan teknologi komputerisasi *Machine Learning* dan memakai metode NBC dan SVM untuk menghitung nilai *accuracy* pada kedua metode tersebut agar dapat mengidentifikasi bahasa sesuai dengan teks yang dimasukkan. Penelitian ini juga melakukan perbandingan dari kedua metode tersebut agar dapat mengetahui metode apa yang lebih efektif dipakai untuk mengidentifikasi bahasa. Dari penelitian yang dilakukan, hasil yang didapatkan dengan menggunakan metode NBC dan SVM adalah keduanya bagus dalam melakukan identifikasi bahasa karena memperoleh nilai *accuracy* di atas 0,90 hanya saja melalui perhitungan *confusion matrix*, metode SVM lebih efektif dengan nilai *accuracy* 0,9634 dibandingkan dengan nilai NBC 0,9378.

Pada penelitian (Yasir & Suraji, 2023), melakukan penelitian tentang perbandingan klasifikasi *naive bayes*, *decision tree*, *random forest* terhadap analisis sentimen kenaikan biaya haji 2023 pada media sosial youtube. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat dengan mengklasifikasikan komentar positif dan negatif di media sosial YouTube terkait kebijakan kenaikan biaya haji tahun 2023. Penelitian ini membandingkan akurasi beberapa metode klasifikasi, seperti *Naïve Bayes*, *Decision Tree*, dan *Random Forest*. Hasil dari penelitian ini menunjukkan bahwa akurasi *Naïve Bayes* mencapai 90%, *Decision Tree* 83%, dan *Random Forest* 87%.