JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 12 No. 3S1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v12i3S1.5149

KLASIFIKASI TINGKAT TUTUR BAHASA SASAK BERBASIS TEKS MENGGUNAKAN NAÏVE BAYES

Lalu Muhamad Alawi Arrozi[1*], Aviv Yuniar Rahman[2], Rangga Pahlevi Putra[3]

[1,2,3] Universitas Widyagama Malang, Jl. Taman Borobudur Indah No.1, Mojolangu, Kec. Lowokwaru, Kota Malang, Jawa Timur 65142

Received: 5 Agustus 2024
Accepted: 5 Oktober 2024
Published: 12 Oktober 2024

Keywords:
Naïve Bayes; Klasifikasi Teks; Tingkat Tutur; Bahasa Sasak; TF-IDF.

Corespondent Email:
alawilalu010601@gmail.com

Abstrak. Tujuan dari penelitian ini adalah untuk menilai efektivitas klasifikasi tingkat tutur bahasa Sasak berbasis teks menggunakan algoritma Naive Bayes dalam mengidentifikasi dan mengkategorikan tingkat tutur bahasa Sasak. Penurunan kesadaran kaum muda mengenai penggunaan "tatakrama" atau tingkat tutur dalam percakapan sehari-hari di Lombok menunjukkan perlunya melestarikan aspek budaya yang penting ini. Tingkat tutur, yang melibatkan sistem kode untuk menyampaikan kesopanan, mencakup kosakata dan aturan leksikal tertentu. Metode yang digunakan dalam penelitian ini adalah Naive Bayes, yang memanfaatkan probabilitas dan statistik untuk klasifikasi teks. Ada dua tahap utama dalam studi ini: pelatihan dan pengujian, dengan pembagian data 70:30. Temuan menunjukkan bahwa model Naive Bayes mencapai F1-score sebesar 84,99%, akurasi 85,08%, presisi 85,12%, dan recall 85,08%. Hasil ini menunjukkan bahwa Naive Bayes adalah metode yang efektif untuk mengklasifikasikan tingkat tutur bahasa Sasak, meskipun hasilnya tidak setinggi beberapa studi sebelumnya. Penelitian ini memberikan kontribusi terhadap pengembangan metode yang lebih efisien dan akurat untuk klasifikasi teks tingkat tutur bahasa Sasak dan menunjukkan perlunya perbaikan dalam pemilihan fitur serta perluasan dataset untuk studi-studi mendatang.

Abstract. *The purpose of this study is to assess the efficacy of text-based Sasak speech level classification utilizing the Naive Bayes algorithm in identifying and categorizing Sasak speech levels. The decline in youth awareness regarding the use of " manners " or speech levels in everyday conversations in Lombok highlights the need to preserve this crucial cultural aspect. Speech levels, which involve a system of codes for conveying politeness, include specific vocabulary and lexical rules. The method employed in this research is Naive Bayes, utilizing probability and statistics for text classification. There are two primary stages to the study: training and testing, with a 70:30 data split. The findings show that the Naive Bayes model attains an F1-score of 84.99%, accuracy of 85.08%, precision of 85.12%, and recall of 85.08%. These findings suggest that Naive Bayes is an effective method for classifying Sasak speech levels, although the results are not as high as some previous studies. This research contributes to the development of more efficient and accurate methods for text-based classification of Sasak speech levels and suggests the need for improvements in feature selection and dataset expansion for future studies.*

1. PENDAHULUAN

Bahasa di sini tidak hanya dianggap sebagai fenomena individual tetapi juga sebagai fenomena sosial. Sebagai manifestasi sosial, penggunaan bahasa dipengaruhi tidak hanya oleh faktor linguistik tetapi juga oleh aspek-aspek seperti status sosial, tingkat pendidikan, usia, ekonomi, jenis kelamin, dan sebagainya. Faktor situasional juga memainkan peran penting dalam menentukan gaya bicara seseorang, di mana bahasa yang digunakan dalam situasi formal dapat berbeda dari yang digunakan dalam situasi santai. Situasi formal cenderung mendorong penutur untuk menggunakan bahasa formal, sementara situasi santai cenderung mempengaruhi penutur untuk menggunakan variasi bahasa informal[1]. Setiap penggunaan bahasa mengikuti norma sosial yang mengendalikan perilaku dan ucapan. Pemilihan kode bahasa penting untuk menyesuaikan diri dengan situasi sosiokultural yang ada[2].

Bahasa Sasak adalah salah satu bahasa daerah di Indonesia yang terdapat di Pulau Lombok, Nusa Tenggara Barat. Bahasa Sasak digunakan oleh masyarakat Pulau Lombok untuk berinteraksi dengan sesama anggota komunitas dalam interaksi sehari-hari[3]. Dalam bahasa Sasak, terdapat tiga tingkat bahasa yang dikenal: Sasak Jamaq (umum) atau aok-ape (ya-apa), Sasak Madya (sopan) atau tiang-inggih (saya-ya), dan Sasak Pembayun (sangat sopan) atau kaji-meran (saya-ya). Ketiga variasi tingkat bahasa Sasak ini memiliki kosakata khas yang berbeda satu sama lain, namun banyak kata yang umum digunakan di semua tingkat, terutama kosakata dari bahasa Jamaq, yang merupakan sumber utama dari seluruh kosakata bahasa Sasak[4].

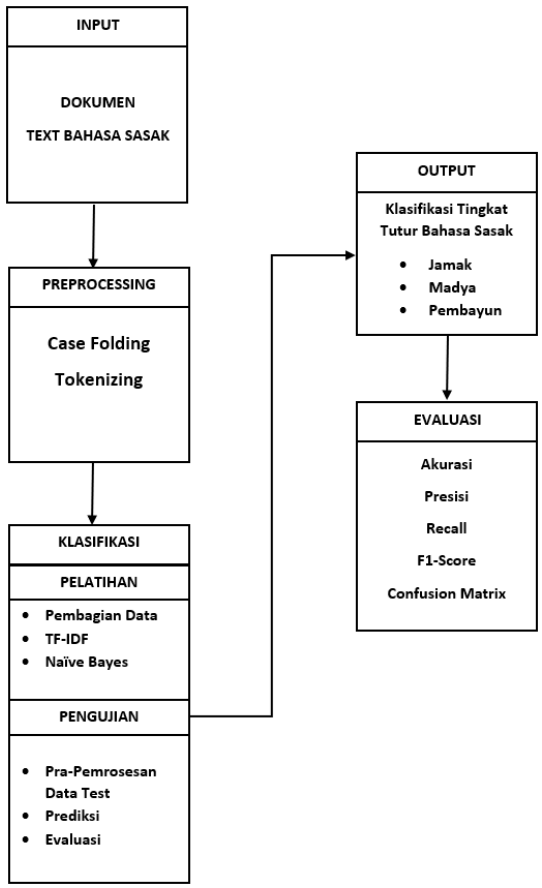
Dalam beberapa tahun terakhir, terutama di Lombok, kesadaran generasi muda untuk menggunakan "tatakrama" atau tingkat bahasa dalam percakapan sehari-hari telah menurun. Jika hal ini terus berlanjut, sebagian dari budaya daerah yang berkontribusi pada budaya nasional Indonesia akan hilang. Tingkat bahasa merujuk pada sistem kode untuk menyampaikan kesopanan, yang melibatkan penggunaan kosakata dan aturan leksikal tertentu[1].

Penelitian tentang klasifikasi tingkat tutur bahasa Sasak berdasarkan teks dengan pendekatan algoritmik belum dilakukan. Namun, terdapat beberapa penelitian terkait tentang klasifikasi teks menggunakan Naive Bayes. Penelitian berjudul "Algoritma Naïve Bayes Untuk Klasifikasi Sumber Belajar Berbasis Teks Pada Mata Pelajaran Produktif di SMK Rumpun Teknologi Informasi dan Komunikasi" oleh Admaja Dwi Herlambang[5]. mencapai kinerja terbaik dengan akurasi 81,48%, sedangkan akurasi terendah adalah 79,63%. Penelitian berjudul "Text Classification With Naïve Bayes" oleh Epri Wahyudi[6] menunjukkan bahwa hasil yang dihitung menggunakan tabel confusion matrix mencapai akurasi maksimum 97%. Penelitian berjudul "Klasifikasi Dialek Bahasa Jawa Menggunakan Metode Naïve Bayes" oleh Grace Angeline, dkk[7]. mencapai kinerja terbaik dengan akurasi 96,97%, presisi 97,53%, dan recall 96,83%. Ini menunjukkan potensi algoritma Naive Bayes untuk klasifikasi teks yang berhasil.

Dalam penelitian ini, penulis menggunakan metode Naive Bayes, yaitu sebuah algoritma untuk klasifikasi teks[8]. Algoritma ini memanfaatkan teknik probabilistik dan statistik yang diusulkan oleh Thomas Bayes, seorang ilmuwan Inggris[9]. Metode Naive Bayes melalui dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap pengujian. Dengan menggunakan algoritma Naive Bayes, diharapkan penelitian ini akan mencapai akurasi yang baik sehingga dapat diimplementasikan dalam klasifikasi tingkat tutur bahasa Sasak berdasarkan teks.

2. METODE PENELITIAN

Gambar 1, menampilkan model yang akan digunakan untuk klasifikasi tingkat tutur bahasa Sasak berbasis teks menggunakan algoritma Naive Bayes.



Gambar 1. Model Klasifikasi Tingkat Tutur Bahasa sasak Berbasis Teks Menggunakan Naïve Bayes

Model klasifikasi tingkat tutur bahasa Sasak berbasis teks menggunakan Naive Bayes dalam penelitian ini dimulai dengan memasukkan dokumen teks bahasa Sasak. Dokumen teks tersebut kemudian diproses melalui tahap praproses, termasuk case folding dan tokenisasi. Selanjutnya, dilakukan pelatihan dan pengujian. Pada tahap pelatihan, data dibagi menjadi data pelatihan dan data uji, kemudian fitur-fitur diekstraksi menggunakan Tf-Idf, dan model Naive Bayes dilatih untuk mengklasifikasikan teks. Pada tahap pengujian, data uji dipraproses terlebih dahulu agar sesuai dengan perlakuan data pelatihan. Setelah pembuatan prediksi menggunakan model untuk mengategorikan data uji, dilakukan penilaian. Hasil dari proses ini adalah tiga tingkat tutur bahasa Sasak, yaitu Jamaq, Madya, dan Pembayun, yang dihasilkan dari klasifikasi teks. Berbagai kriteria digunakan dalam evaluasi keseluruhan, seperti recall, f1-score, akurasi, presisi, dan confusion matrix.

2.1. Input

Data diperoleh melalui observasi masyarakat dan wawancara dengan penutur asli bahasa Sasak dari berbagai latar belakang sosial dan usia. Wawancara ini dilakukan untuk mengumpulkan contoh penggunaan tingkat tutur dalam konteks sehari-hari. Total data yang dikumpulkan terdiri dari 1.050 kalimat yang diklasifikasikan ke dalam tiga tingkat tutur utama dalam bahasa Sasak: Jamaq, Madya, dan Pembayun. Setiap kalimat kemudian diberi label sesuai dengan tingkat tutur berdasarkan penilaian dari ahli bahasa Sasak dan penutur asli. Proses pengumpulan data mematuhi etika penelitian, termasuk memperoleh persetujuan dari responden wawancara dan memastikan anonimitas data yang dikumpulkan dari sumber online. Dengan demikian, data yang digunakan dalam penelitian ini diharapkan representatif dan berkualitas tinggi untuk tujuan klasifikasi tingkat tutur bahasa Sasak.

2.2. Preprocessing

Mengubah semua huruf dalam sebuah kalimat menjadi huruf kecil dikenal sebagai case folding. Proses ini bertujuan untuk menstandarkan format teks agar analisis menjadi lebih konsisten[8]. Tabel 1 menyajikan dataset yang telah mengalami case folding.

Table 1. Hasil Case Folding

Sebelum Case Folding	Setelah Case Folding
Side Wah Toak	side wah toak
Kanggo Aku Beketuan	kanggo aku beketuan
Tiang Sampun Belek	tiang sampun belek

Setiap kalimat dipecah menjadi unit yang lebih kecil yang disebut token. Tokenisasi membantu dalam menganalisis frekuensi kata dan pola bahasa dalam teks[10]. Tabel 2 menyajikan dataset yang telah mengalami tokenisasi.

Table 2. Hasil Tokenisasi

Sebelum Tokenisasi	Setelah Tokenisasi
side wah toak	side wah toak

kanggo aku beketuan	kanggo aku beketuan
tiang sampun belek	tiang sampun belek

2.3. Klasifikasi

Tahap ini adalah aspek yang paling krusial dalam proses penelitian, yaitu pengumpulan data yang diperlukan. Data terdiri dari kalimat teks dari tiga tingkat tutur bahasa Sasak, yang diperoleh melalui percakapan masyarakat. Sebanyak 1.050 kalimat dikumpulkan, dan data akan dibagi menjadi dua set: data pelatihan dan data uji, dengan rasio pemisahan berkisar antara 10:90 hingga 90:10.

Karakteristik utama diperoleh dari data melalui teknik TF-IDF (Term Frequency-Inverse Document Frequency). Metode ini memberikan bobot pada setiap kata dalam sebuah dokumen sesuai dengan frekuensinya dalam dokumen tersebut dan prevalensinya di seluruh dokumen. Metode ini membantu dalam mengidentifikasi kata-kata yang penting secara kontekstual dan meminimalkan dampak kata-kata yang umum[11].

Algoritma Naive Bayes adalah metode klasifikasi probabilistik yang didasarkan pada Teorema Bayes. Teorema Bayes pada dasarnya menghitung probabilitas bersyarat secara terbalik, memperbarui keyakinan tentang suatu peristiwa setelah mengamati bukti atau informasi baru. Algoritma Naive Bayes menerapkan Teorema Bayes untuk klasifikasi dengan asumsi bahwa setiap fitur dalam data bersifat independen terhadap fitur lainnya, meskipun asumsi ini mungkin tidak selalu sepenuhnya akurat dalam praktiknya. Algoritma Naive Bayes sering digunakan dalam tugas klasifikasi teks dan dokumen. Model ini efektif untuk tugas klasifikasi teks, di mana setiap atribut independen diberikan label kelas. Keunggulan Naive Bayes termasuk kemudahan penggunaan, kecepatan, dan akurasi yang tinggi[12].

2.4. Output

Dalam klasifikasi tingkat tutur bahasa Sasak, sub-output dikategorikan ke dalam tiga tingkat yang berbeda: Jamaq, Madya, dan Pembayun. Setiap tingkat mewakili derajat kesopanan dan formalitas yang berbeda dalam bahasa Sasak, dengan Jamaq sebagai tingkat yang paling umum, Madya sebagai tingkat sopan, dan Pembayun sebagai tingkat yang paling formal. Model Naive Bayes memproses tingkat-tingkat ini dengan menganalisis fitur teks dan memberikan label kelas berdasarkan data yang telah dilatih, sehingga sistem dapat mengklasifikasikan input teks baru ke dalam salah satu dari ketiga kategori ini. Efektivitas klasifikasi ini dievaluasi dengan mengukur akurasi, presisi, recall, dan f1-score dari model dalam mengidentifikasi setiap tingkat tutur dengan benar.

2.5. Evaluasi

Evaluasi model klasifikasi berbasis teks untuk tingkat tutur bahasa Sasak menggunakan Naive Bayes melibatkan beberapa metrik: recall, F1-score, akurasi, presisi, dan confusion matrix untuk mengetahui seberapa baik kinerja pemodelan sebelumnya[13].

Akurasi, seperti yang disajikan dalam persamaan (1), mengukur tingkat kebenaran keseluruhan model dengan menghitung rasio antara jumlah prediksi yang benar dengan jumlah total instance:

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \quad (1)$$

Akurasi yang tinggi menunjukkan bahwa model berfungsi dengan baik di semua kelas.

Presisi, seperti yang disajikan dalam persamaan (2), menentukan rasio kasus positif yang benar terhadap jumlah total kasus positif yang benar dan kasus positif yang salah untuk menilai akurasi prediksi positif.

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

Presisi yang tinggi berarti bahwa ketika model memprediksi tingkat tutur tertentu, kemungkinan besar prediksinya benar.

Recall (atau Sensitivitas), seperti yang disajikan dalam persamaan (3), mengukur kemampuan model untuk mengidentifikasi semua instance relevan dari suatu kelas dengan menghitung rasio antara jumlah instance positif yang benar dengan jumlah total instance positif yang benar dan instance positif yang salah.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Recall yang tinggi menunjukkan bahwa model berhasil mengidentifikasi sebagian besar instance dari tingkat tutur tertentu.

F1-Score, seperti yang disajikan dalam persamaan (4), menggabungkan presisi dan recall menjadi satu metrik dengan menghitung rata-rata harmonisnya[14]. Ini menyeimbangkan trade-off antara presisi dan recall.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

F1-score yang tinggi mencerminkan keseimbangan yang baik antara presisi dan recall.

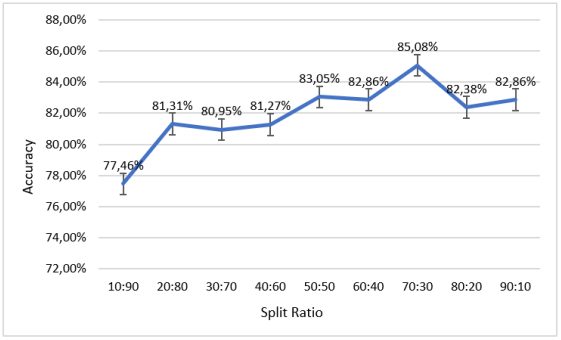
Confusion Matrix memberikan analisis menyeluruh tentang kinerja model dengan menunjukkan jumlah true positives, false positives, true negatives, dan false negatives untuk setiap kelas[15]. Ini membantu dalam memahami kelas mana yang sering tertukar dengan kelas lainnya dan memberikan wawasan tentang kinerja model di berbagai tingkat tutur.

Secara keseluruhan, metrik-metrik ini memberikan penilaian yang komprehensif terhadap kinerja model Naive Bayes dalam mengklasifikasikan tingkat tutur bahasa Sasak, memastikan bahwa klasifikasi tersebut akurat dan dapat diandalkan.

3. HASIL DAN PEMBAHASAN

3.1. Perbandingan Split Rasio dengan Akurasi

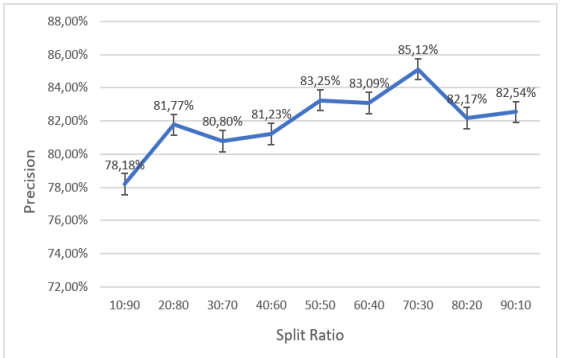
Model Naive Bayes menunjukkan kinerja yang cukup konsisten di berbagai rasio data pelatihan dan pengujian, dengan akurasi berkisar antara 77,46% hingga 85,08%. Akurasi tertinggi dicapai dengan rasio data pelatihan 70% dan data pengujian 30% (85,08%), menunjukkan bahwa model berfungsi paling baik dengan distribusi data ini. Dengan 70% data digunakan untuk pelatihan, model memiliki informasi yang cukup untuk mempelajari pola dalam data, sehingga meningkatkan akurasinya dalam memprediksi hasil pada data pengujian. Rasio ini juga memberikan keseimbangan yang baik antara data pelatihan dan pengujian, memungkinkan model untuk menggeneralisasi secara efektif pada data baru. Risiko overfitting dan underfitting diminimalkan dengan rasio ini, dan kompleksitas model sesuai. Variasi dalam akurasi di berbagai rasio pemisahan data mencerminkan keseimbangan antara kemampuan model untuk mempelajari pola dari data pelatihan dan kemampuannya untuk diuji secara representatif dengan data pengujian. Perbandingan akurasi untuk setiap rasio pemisahan dapat dilihat pada Gambar 2.



Gambar 2. Pebandingan Split Rasio dengan Akurasi

3.2. Perbandingan Split Rasio dengan Presisi

Secara umum, presisi model berkisar antara 80,80% hingga 85,12%. Rasio data pelatihan 70% dan rasio data pengujian 30% menghasilkan nilai presisi maksimum sebesar 85,12%. Perbandingan presisi untuk setiap rasio pemisahan dapat dilihat pada Gambar 3.

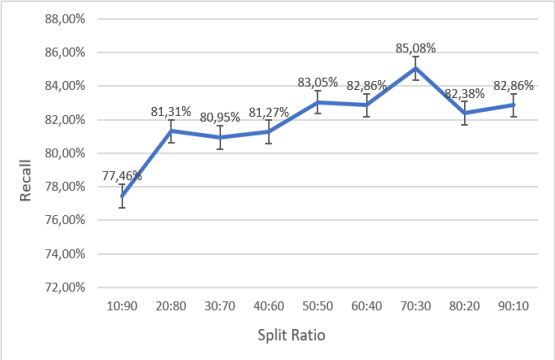


Gambar 3. Perbandingan Split Rasio dengan Presisi

3.3. Perbandingan Split Rasio dengan Recall

Recall model bervariasi antara 80,95% hingga 85,08%, dengan nilai tertinggi 85,08% dicapai pada rasio data pelatihan 70% dan data pengujian 30%. Ini menunjukkan bahwa model

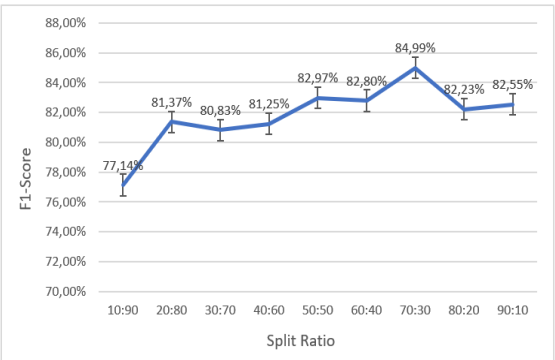
paling efektif dalam mengidentifikasi instance relevan ketika menggunakan rasio pemisahan ini. Kemampuan model untuk mengingat instance relevan sangat penting untuk memastikan bahwa model bekerja dengan baik di berbagai tingkat tutur. Perbandingan rinci recall untuk setiap rasio pemisahan diilustrasikan pada Gambar 4.



Gambar 4. Perbandingan Split Rasio dengan Recall

3.4. Perbandingan Split Rasio dengan F1-score

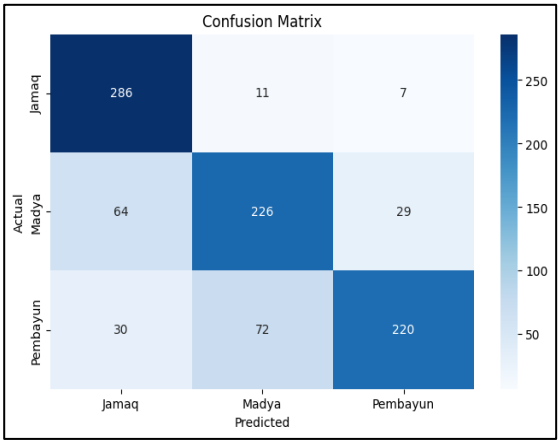
Nilai F1-score mencerminkan keseimbangan antara presisi dan recall, berkisar antara 80,83% hingga 84,99%. F1-score tertinggi sebesar 84,99% dicapai pada rasio data pelatihan 70% dan data pengujian 30%. Ini menunjukkan bahwa model mempertahankan keseimbangan yang baik antara presisi dan recall pada rasio ini. Kinerja F1-score di berbagai rasio pemisahan diilustrasikan pada Gambar 5.



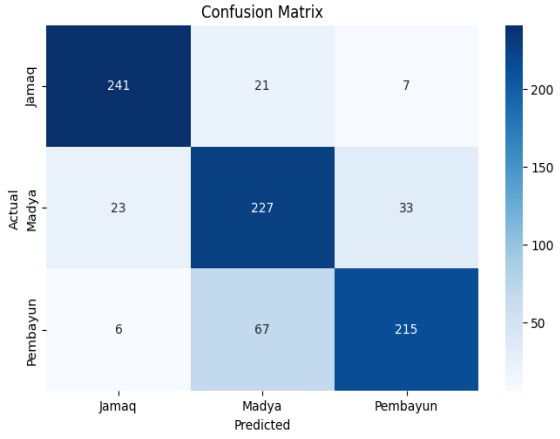
Gambar 5. Perbandingan Split Rasio dengan F1-score

3.5. Perbandingan Confusion Matrix

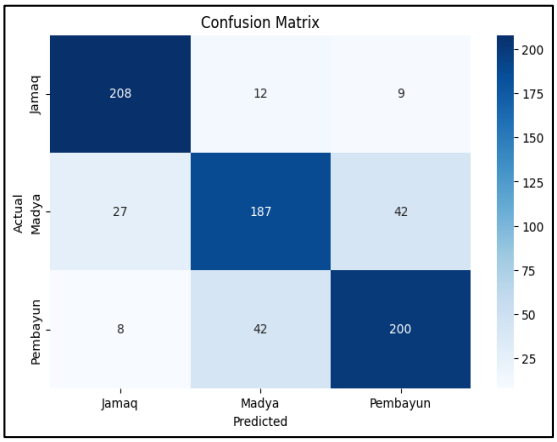
Gambar 6 hingga Gambar 14 menyajikan hasil confusion matrix untuk model klasifikasi dengan berbagai rasio pemisahan antara data pelatihan dan data uji, berkisar dari 10:90 hingga 90:10. Gambar 6 menunjukkan hasil confusion matrix untuk rasio data pelatihan 10% dan data uji 90%, memberikan wawasan tentang bagaimana model berfungsi pada data uji dengan proporsi data pelatihan yang kecil. Gambar 7 menampilkan hasil untuk rasio 20:80, Gambar 8 untuk rasio 30:70, Gambar 9 untuk rasio 40:60, Gambar 10 untuk rasio 50:50, Gambar 11 untuk rasio 60:40, Gambar 12 untuk rasio 70:30, Gambar 13 untuk rasio 80:20, dan Gambar 14 untuk rasio 90:10. Setiap gambar menggambarkan perbedaan dalam kinerja klasifikasi berdasarkan perubahan rasio data pelatihan dan data uji serta memberikan wawasan tentang bagaimana proporsi data pelatihan mempengaruhi kemampuan model dalam memprediksi kelas yang berbeda.



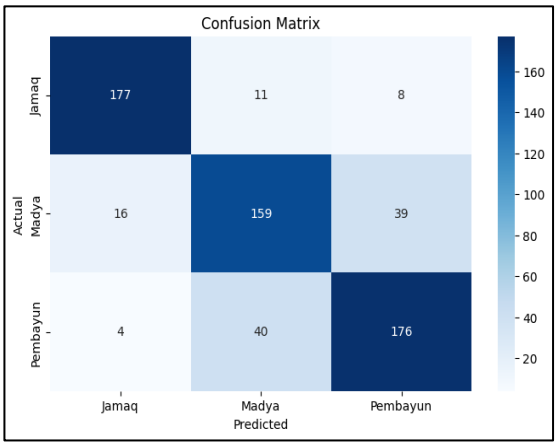
Gambar 6. Perbandingan Confusion Matrix dengan Split Rasio 10:90



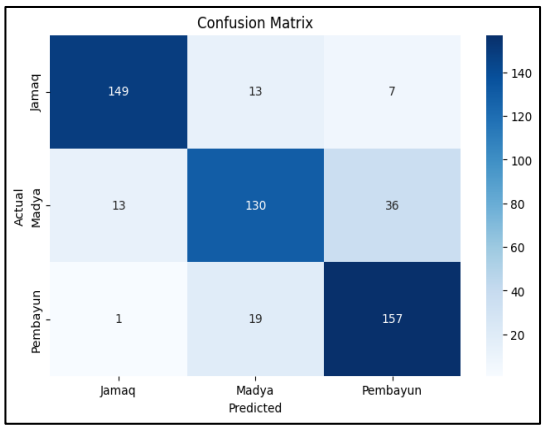
Gambar 7. Perbandingan Confusion Matrix dengan Split Rasio 20:80



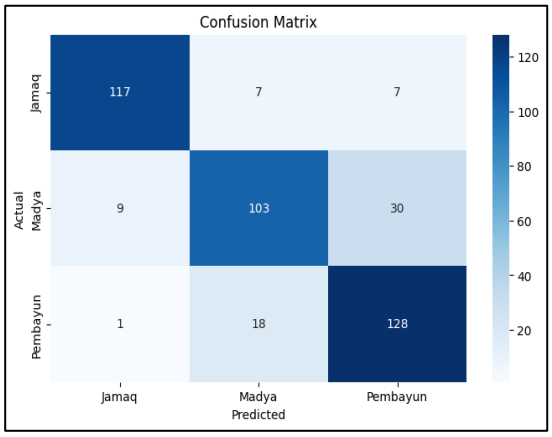
Gambar 8. Perbandingan Confusion Matrix dengan Split Rasio 30:70



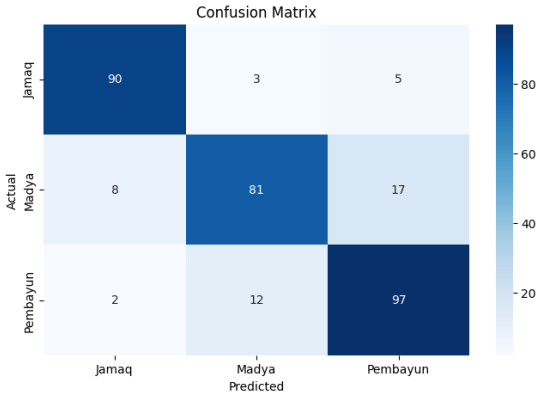
Gambar 9. Perbandingan Confusion Matrix dengan Split Rasio 40:60



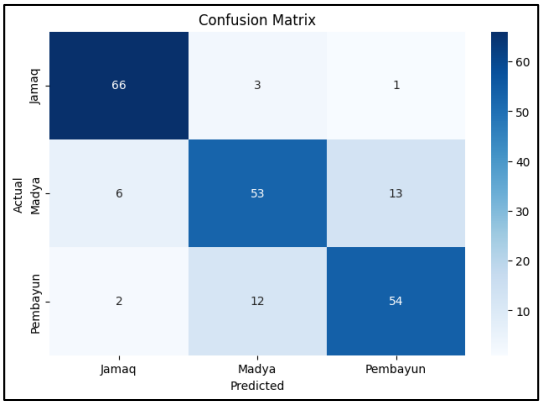
Gambar 10. Perbandingan Confusion Matrix dengan Split Rasio 50:50



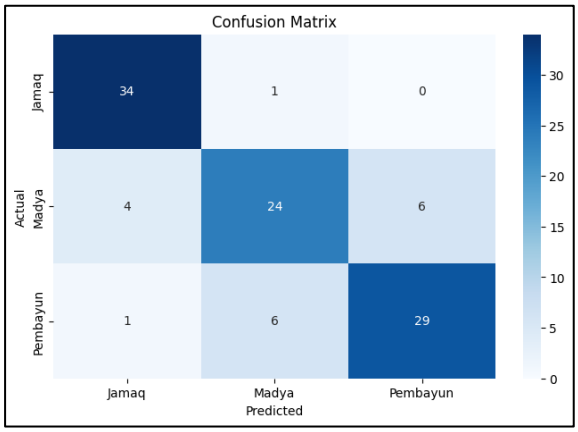
Gambar 11. Gambar 10. Perbandingan Confusion Matrix dengan Split Rasio 60:40



Gambar 12. Perbandingan Confusion Matrix dengan Split Rasio 70:30



Gambar 13. Perbandingan Confusion Matrix dengan Split Rasio 80:20



Gambar 14. Perbandingan Confusion Matrix dengan Split Rasio 90:10

Hasil dari confusion matrix pada berbagai rasio pemisahan data pelatihan dan data uji menunjukkan bahwa seiring dengan meningkatnya proporsi data pelatihan (misalnya, dari 10% hingga 90%), model klasifikasi semakin baik dalam memprediksi kelas-kelas umum seperti Jamaq dan Pembayun. Namun, jika rasio data pelatihan terlalu tinggi (misalnya, 80% atau 90%), model cenderung mengalami overfitting, yang mengakibatkan penurunan kinerja pada data uji karena model menjadi terlalu spesifik terhadap data pelatihan. Sebaliknya, dengan rasio data pelatihan yang lebih rendah (misalnya, 10% hingga 30%), model mungkin tidak memiliki informasi yang cukup untuk belajar secara optimal, yang mengakibatkan banyak kesalahan prediksi. Oleh karena itu, rasio pemisahan data pelatihan sekitar 30% hingga 50% umumnya memberikan keseimbangan terbaik antara pelatihan dan evaluasi, menghasilkan kinerja yang lebih stabil pada data uji.

3.6. Diskusi Hasil Penelitian Sebelumnya dengan Temuan Penelitian yang Diusulkan

Tabel 3 menyajikan perbandingan temuan penelitian sebelumnya dengan penelitian yang diusulkan dalam studi ini. Tabel tersebut membandingkan berbagai aspek yang menjadi fokus analisis dalam studi-studi sebelumnya dengan temuan dan metodologi yang digunakan dalam penelitian ini. Secara spesifik, tabel mencakup metrik seperti akurasi (A), presisi (P), recall (R), dan F1 score (F). Metrik-metrik ini memberikan evaluasi yang komprehensif tentang kinerja berbagai model dan metodologi, memungkinkan perbandingan mendalam mengenai efektivitasnya.

Tabel 3. Diskusi Penelitian Terdahulu

Penelitian	Metode	Evaluasi			
		A	P	R	F
[7]	Naïve Bayes	96	97	96	-
[6]	Naïve Bayes	97	-	-	-
[5]	Naïve Bayes	81	-	-	-
Penelitian Kami	Naïve Bayes	85	85	85	84

Penelitian yang dilakukan oleh Angeline, Wibawa, dan Pujianto[7] menggunakan metode Naïve Bayes dan menunjukkan bahwa akurasi yang dicapai adalah 96,97%, presisi 97,53%, dan recall 96,83%. Hasil ini menunjukkan kinerja yang sangat baik dalam mengklasifikasikan dialek Jawa, melebihi temuan dari studi saat ini. Ini menunjukkan bahwa meskipun Naïve Bayes secara umum efektif, kinerjanya dapat bervariasi tergantung pada data dan konteks yang digunakan. Sebaliknya, Wahyudi[6] menggunakan Naïve Bayes Classifier (NBC) dan melaporkan akurasi sebesar 97%. Meskipun presisi, recall, dan F1-Score tidak dilaporkan, akurasi yang tinggi ini lebih besar dibandingkan dengan studi saat ini, menunjukkan bahwa Naïve Bayes dapat mencapai akurasi yang sangat tinggi, meskipun hasilnya tetap bisa bervariasi berdasarkan faktor-faktor yang berbeda.

Di sisi lain, Hadi Wijoyo dan Dwi Herlambang[5] menerapkan metode Naïve Bayes dan melaporkan akurasi sebesar 81,48%. Hasil ini lebih rendah dibandingkan dengan studi saat ini, tetapi tetap menunjukkan bahwa Naïve Bayes tetap efektif di berbagai aplikasi. Studi saat ini, yang menggunakan Naïve Bayes, menemukan akurasi sebesar 85,08%, presisi 85,12%, recall 85,08%, dan F1-Score 84,99%. Meskipun hasil ini menunjukkan kinerja yang baik, hasil tersebut tidak mencapai tingkat yang dilaporkan oleh Angeline, Wibawa, dan Pujianto atau Wahyudi. Namun, hasilnya lebih tinggi daripada yang ditemukan oleh Hadi Wijoyo dan Dwi Herlambang, menunjukkan bahwa kinerja Naïve Bayes dapat bervariasi tergantung pada data dan konteks yang digunakan.

4. KESIMPULAN DAN SARAN

4.1. Kesimpulan

Berdasarkan analisis dan hasil evaluasi klasifikasi tingkat tutur bahasa Sasak berbasis teks menggunakan algoritma Naive Bayes dengan fitur TF-IDF, beberapa kesimpulan kunci dapat diambil:

a. Model Naive Bayes mencapai akurasi sebesar 85,08% dalam mengklasifikasikan tingkat tutur bahasa Sasak ke dalam tiga kelas: Jamaq, Madya, dan Pembayun pada rasio data pelatihan 70% dan data pengujian 30%.
b. Model ini menunjukkan kinerja yang sangat baik dalam mengklasifikasikan kelas Jamaq, cukup efektif untuk kelas Pembayun, tetapi menunjukkan kelemahan dalam mengklasifikasikan kelas Madya.
c. Rasio data pelatihan dan pengujian yang optimal adalah 70% dan 30%, di mana model mencapai akurasi dan F1-score tertinggi.
d. Dengan rasio pelatihan 70% dan pengujian 30%, model Naive Bayes mampu memberikan kinerja terbaik, menyeimbangkan antara akurasi dan F1-score, serta menunjukkan kemampuan yang baik dalam mengidentifikasi kelas-kelas tutur bahasa Sasak.

4.2. Saran

Untuk meningkatkan sistem klasifikasi tingkat tutur bahasa Sasak berbasis teks menggunakan algoritma Naive Bayes, beberapa perubahan diperlukan untuk mencapai hasil yang lebih baik daripada penelitian ini. Penelitian lanjutan diperlukan untuk meningkatkan kinerja model pada kelas Madya. Selain itu, meningkatkan jumlah data sangat penting untuk mengoptimalkan kinerja algoritma. Perbaikan dalam pemilihan fitur juga diperlukan, dan mempertimbangkan fitur tambahan di luar TF-IDF mungkin bermanfaat. Penelitian mendatang sebaiknya bertujuan untuk memperluas dataset dengan menggabungkan lebih banyak variasi dalam konteks sosial dan budaya bahasa Sasak. Menambahkan data dari berbagai daerah di Lombok dapat memberikan pandangan yang lebih komprehensif tentang variasi dalam tingkat tutur bahasa Sasak.

UCAPAN TERIMA KASIH

Peneliti ingin mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Lalu Mawardi Zain, QH., SS., S.Pd., atas bantuannya dalam memvalidasi data yang dikumpulkan. Dukungan dan keahliannya telah sangat penting dalam memastikan kualitas dan akurasi data penelitian ini. Selain itu, peneliti juga ingin menyampaikan penghargaan yang mendalam kepada pembimbing jurnal, yang keahlian dan dedikasinya telah berkontribusi secara signifikan terhadap pengembangan dan kesuksesan studi ini. Terima kasih atas bimbingan dan dukungan yang terus menerus. Peneliti juga ingin mengucapkan terima kasih kepada lembaga-lembaga terkait yang telah menyediakan dukungan dan fasilitas yang diperlukan untuk melaksanakan penelitian ini. Dukungan dari berbagai pihak telah memungkinkan studi ini berjalan dengan lancar dan mencapai tujuannya.

DAFTAR PUSTAKA

[1] L. C. U. Buana, “Tingkat Tutur dalam Bahasa Sasak Desa Pagutan Kabupaten Lombok Tengah,” *J. Bastrindo*, 2023.
[2] M. D. B. Akastangga, “Dialek Sebagai Identitas Masyarakat Bahasa di Pulau Lombok,” *Int. Semin. Austronesian Lang. Lit. IX*, no. September, pp. 139–146, 2021.
[3] L. Hakim, “Sapaan Kekerabatan Bahasa Sasak Di Desa Beraim, Kecamatan Praya Tengah, Lombok Tengah,”,” *MABASAN*, vol. 14, no. 2, pp. 329–340, Dec. 2020, doi: 10.26499/mab.v14i2.426.
[4] H. D. Ikawati and Z. Anwar, “Pengembangan Sumber Belajar Muatan Lokal Bahasa Sasak Halus,” *J. Sci. …*, vol. 2, no. 11, pp. 582–590, 2021.
[5] S. Hadi Wijoyo and A. Dwi Herlambang, “Algoritma Naïve Bayes Untuk Klasifikasi Sumber Belajar Naïve Bayes Algorithm for Text Based Learning Resources Classification in Productive Subject At Information,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 4, pp. 431–436, 2019, doi: 10.25126/jtiik.201961323.
[6] E. Wahyudi, “Text Classification With Naïve Bayes,” *Teknologipintar.org*, vol. 2, no. 4, 2022.
[7] G. Angeline, A. P. Wibawa, and U. Pujianto, “Klasifikasi Dialek Bahasa Jawa Menggunakan Metode Naive Bayes,” *J. Mnemon.*, vol. 5, no. 2,

pp. 103–110, 2022, doi: 10.36040/mnemonic.v5i2.4748.

[8] S. Supriyatna and E. Fahrudin, "Pemanfaatan Algoritma Text Mining Dalam Menemukan Pola Risiko Bencana Sebagai Pengetahuan Kebencanaan Dari Dokumen Kajian Risiko Bencana (Krb) 1*," *J. Inform. Utama*, vol. 2, no. 1, pp. 35–42, 2024, [Online]. Available: https://doi.org/10.55903/jitu.v2i1.xx

[9] R. Rahayu, "Algoritma Naive Bayes," *Res. Artic. · December 2023*, Dec. 2023.

[10] N. Pittaras, G. Giannakopoulos, G. Papadakis, and V. Karkaletsis, *Text classification with semantically enriched word embeddings*, vol. 27, no. 4. 2021. doi: 10.1017/S1351324920000170.

[11] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[12] E. Indrayuni, S. Sistem, I. A. Kampus, and K. Bogor, "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes," *J. KHATULISTIWA Inform.*, vol. VII, no. 1, 2019.

[13] L. Nursinggah, T. Mufizar, and U. Perjuangan, "Analisis Sentimen Pengguna Aplikasi X Terhadap Program Makan Siang Gratis Dengan Metode Naïve Bayes Classifier," *J. Inform. dan Tek. Elektro Ter.*, vol. 12, no. 3, 2024.

[14] Yudi Widhiyasana, Transmissia Semiawan, Ilham Gibran Achmad Mudzakir, and Muhammad Randi Noor, "Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 4, pp. 354–361, 2021, doi: 10.22146/jnteti.v10i4.2438.

[15] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform.*, vol. 5, no. 2, pp. 697–711, 2021.