Journal of Computer Science and Technology
JCS-TECH Vol.5, No.1, Mei 2025, PP. 24-28
e-ISSN 2808-9677, p-ISSN 2809-1140

KLASIFIKASI JUDUL SKRIPSI MAHASISWA BERDASARKAN KONSENTRASI PROGRAM STUDI MENGGUNAKAN ALGORITMA NAIVE BAYES

Eva Yumami[1], Miftahul Jannah[2], Khelvin Ovella Putra[3]
[1,2] Politeknik Negeri Bengkalis, Bengkalis
[3]Institut Teknologi Mitra Gama, Duri
[1]evayumami@polbeng.ac.id, [2]miiftahuljannah32@gmail.com, [3]khelvinovela@gmail.com

Abstrak

Penelitian ini membahas pengembangan model klasifikasi otomatis untuk judul skripsi mahasiswa berdasarkan konsentrasi program studi dengan menggunakan algoritma Naive Bayes. Latar belakang dari penelitian ini adalah meningkatnya jumlah judul skripsi yang dihasilkan setiap tahun yang menyebabkan kompleksitas dalam proses klasifikasi dan pengelolaan secara manual. Algoritma Naive Bayes dipilih karena sifatnya yang sederhana, efisien, dan cocok untuk klasifikasi teks. Penelitian ini menggunakan data judul skripsi dari mahasiswa Program Studi D4 Rekayasa Perangkat Lunak Jurusan Teknik Informatika Politeknik Negeri Bengkalis dengan enam konsentrasi keahlian, yaitu Software Engineer, Mobile Developer, Full Stack Developer, UI/UX Designer, Software Quality Assurance Engineer, dan Technopreneur. Data yang digunakan mengalami serangkaian proses pra-pemrosesan seperti tokenisasi, stopword removal, dan stemming. Model dilatih dan diuji menggunakan pendekatan pembagian data latih dan uji dengan evaluasi performa menggunakan metrik akurasi, precision, recall, dan f1-score. Hasil penelitian menunjukkan bahwa algoritma Naive Bayes dapat mengklasifikasikan judul skripsi ke dalam konsentrasi yang sesuai dengan tingkat akurasi mencapai 65%. Penelitian ini berkontribusi dalam mendukung efisiensi pengelolaan administrasi akademik dan dapat dijadikan landasan bagi pengembangan sistem klasifikasi berbasis AI di lingkungan pendidikan tinggi.

Kata Kunci: Naive Bayes, Klasifikasi Teks, Judul Skripsi, Konsentrasi Program Studi.

Abstract

Abstract: This study discusses the development of an automatic classification model for student thesis titles based on study program concentrations using the Naive Bayes algorithm. The background of this research is the increasing number of thesis titles produced each year, which complicates the classification and management process if done manually. The Naive Bayes algorithm was chosen for its simplicity, efficiency, and suitability for text classification tasks. The dataset comprises thesis titles from students of the D4 Software Engineering Program with six areas of concentration: Software Engineer, Mobile Developer, Full Stack Developer, UI/UX Designer, Software Quality Assurance Engineer, and Technopreneur. The data underwent several preprocessing stages including tokenization, stopword removal, and stemming. The model was trained and tested using a train-test split approach and evaluated using accuracy, precision, recall, and f1-score metrics. The results indicate that the Naive Bayes algorithm can classify thesis titles into their appropriate concentrations with an accuracy of 65%. This research contributes to improving the efficiency of academic administration management and serves as a foundation for developing AI-based classification systems in higher education.

Keywords: Naive Bayes, Text Classification, Thesis Title, Study Program Concentration.

1. Pendahuluan

Dalam era digitalisasi pendidikan tinggi, jumlah judul skripsi yang dihasilkan mahasiswa setiap tahunnya terus meningkat secara eksponensial, seiring dengan bertambahnya jumlah mahasiswa dan diversifikasi konsentrasi program studi. Fenomena ini menimbulkan tantangan baru dalam pengelolaan, pengarsipan, serta pemetaan judul skripsi agar sesuai dengan konsentrasi keilmuan yang ada.

Permasalahan klasifikasi judul skripsi berdasarkan konsentrasi program studi semakin kompleks seiring dengan pertumbuhan jumlah mahasiswa dan keragaman topik penelitian di perguruan tinggi. Setiap tahun, institusi pendidikan tinggi menghadapi tantangan dalam mengelola ribuan judul skripsi yang harus dipetakan secara tepat ke

dalam konsentrasi program studi yang relevan. Proses klasifikasi manual tidak hanya memakan waktu dan sumber daya, tetapi juga rentan terhadap subjektivitas serta inkonsistensi, terutama pada program studi dengan banyak konsentrasi dan jumlah mahasiswa yang besar (Ilmiah & Komputer, 2024)

Perkembangan teknologi data mining dan machine learning menawarkan solusi otomatisasi klasifikasi judul skripsi berbasis algoritma, salah satunya adalah Naive Bayes Classifier. Algoritma ini dikenal sederhana, efisien, dan mampu menangani data berukuran besar, sehingga banyak digunakan dalam berbagai aplikasi klasifikasi teks, termasuk pengelompokan dokumen dan analisis sentiment (Suppa, 2023)

Kebutuhan akan sistem klasifikasi otomatis berbasis kecerdasan buatan menjadi semakin mendesak untuk mendukung efisiensi administrasi akademik dan memastikan relevansi topik penelitian mahasiswa dengan kompetensi program studi. pengembangan sistem klasifikasi otomatis berbasis kecerdasan buatan, khususnya algoritma Naive Bayes, semakin tinggi untuk mendukung efisiensi administrasi akademik dan memastikan relevansi topik penelitian mahasiswa. Naive Bayes dikenal sebagai algoritma klasifikasi probabilistik yang sederhana namun efektif dalam menangani data teks, termasuk dalam aplikasi klasifikasi judul skripsi (Ilmiah & Komputer, 2024)

Namun, penerapan Naive Bayes pada klasifikasi judul skripsi menghadapi tantangan tersendiri, seperti tingginya dimensi fitur akibat variasi kata dalam judul, serta asumsi independensi antar fitur yang seringkali tidak terpenuhi dalam data nyata, sehingga dapat menurunkan akurasi klasifikasi (Thesis, 2012) Selain itu, perbandingan dengan algoritma lain seperti Decision Tree, K-NN, dan Neural Network menunjukkan bahwa Naive Bayes memiliki keunggulan dalam kecepatan dan kemudahan implementasi, namun terkadang kalah dalam hal akurasi pada data yang kompleks atau tidak memenuhi asumsi independensi fitur (Nuraeni et al., 2021)(Mubarak et al., 2024)

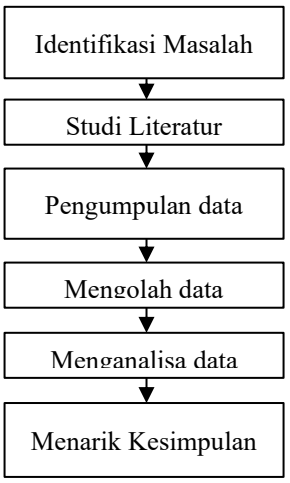
Namun demikian, terdapat kesenjangan penelitian (research gap) yang signifikan. Studi-studi sebelumnya umumnya hanya berfokus pada satu program studi atau konsentrasi tertentu, sehingga model yang dihasilkan kurang mampu melakukan generalisasi pada data dari konsentrasi yang berbeda (Dhuhita et al., 2022) Penelitian sebelumnya yang dilakukan oleh Sukriadi (2023) menerapkan text mining dan algoritma Naive Bayes untuk klasifikasi judul skripsi mahasiswa secara otomatis sesuai konsentrasi. Penelitian ini bertujuan memudahkan program studi dalam memonitoring kesesuaian roadmap penelitian mahasiswa. Hasil implementasi menunjukkan sistem berjalan baik dengan tingkat kesesuaian klasifikasi sebesar 65% untuk kategori sesuai konsentrasi (Sukriadi et al., 2023) dan peneliti berikutnya oleh Denny Kurniadi (2020) mengeksplorasi penerapan metode Naive Bayes dalam memprediksi penentuan topik tugas akhir mahasiswa berdasarkan bidang minat. Penelitian ini membuktikan bahwa metode Naive Bayes mampu memberikan prediksi yang akurat terkait penentuan topik tugas akhir, dengan nilai akurasi, presisi, dan recall yang signifikan pada masing-masing kategori bidang minat (Kurniadi, 2024)

Oleh karena itu, penelitian ini menjadi penting untuk mengkaji lebih lanjut bagaimana optimalisasi algoritma Naive Bayes, khususnya dalam konteks klasifikasi judul skripsi mahasiswa, dapat meningkatkan akurasi dan efisiensi proses klasifikasi, serta mengatasi keterbatasan yang ada pada penelitian-penelitian sebelumnya.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi judul skripsi berbasis Naive Bayes yang dioptimalkan untuk multi-konsentrasi program studi dengan validasi yang ketat dan teknik praproses yang sesuai dengan karakteristik bahasa Indonesia.

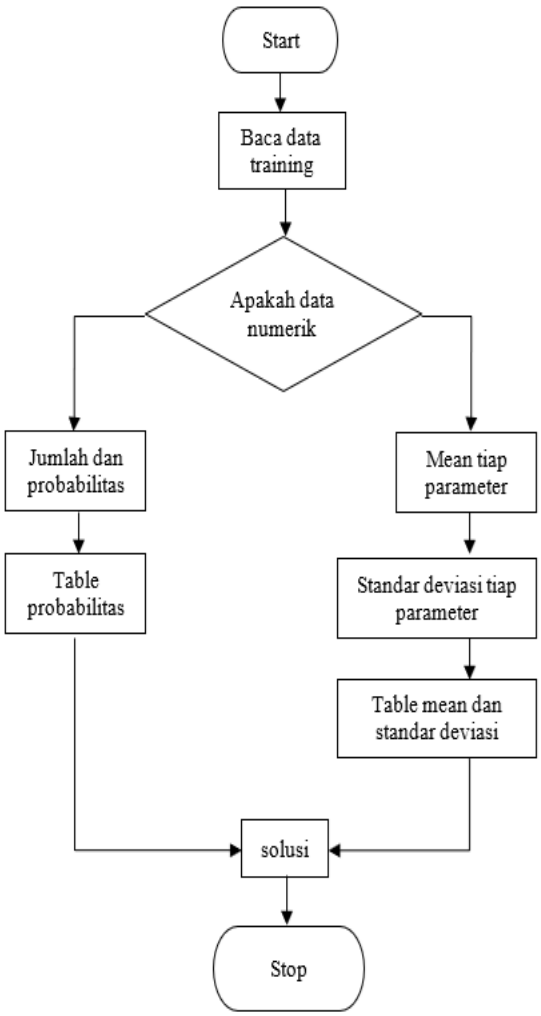
2. Metode Penelitian

Kerangka penelitian agar penelitian dapat diselesaikan secara teratur, terarah dan sistematis maka diperlukan tahap-tahap. Adapun kerangka kerja penelitian dapat digambarkan sebagai berikut :



Gambar 1. Tahapan Penelitian

Langkah awal dilakukan identifikasi masalah yang ada dan dilakukan analisis. Selanjutnya mempelejari literatur berdasarkan penelitian terdahulu sehingga dapat dijadikan dasar dalam penelitian, lalu mengumpulkan data sebanyak 100 judul skripsi mahasiswa prodi D-4 Rekayasa Perangkat Lunak Jurusan Teknik Informatika Politeknik Negeri Bengkalis dan 6 konsentrasi program studi yaitu Software Engineer, Mobile Developer, Full Stack Developer, UI/UX Designer, Software Quality Assurance Engineer, dan Technopreneur. Setelah data terkumpul dilanjutkan mengolah data dengan beberapa tahap diantaranya: data cleaning, data integrasi, data seleksi, data transformasi dan dilanjutkan dengan proses data mining dengan metode naïve bayes seperti gambar dibawah ini.



Gambar 2. Alur metode naïve bayes

3. Hasil Penelitian dan Pembahasan

Bagian ini menyajikan hasil penerapan algoritma Naive Bayes untuk mengklasifikasikan judul skripsi mahasiswa sesuai dengan konsentrasi program studi. Setiap tahapan dalam proses data mining, mulai dari pra-pemrosesan data, pelatihan dan pengujian model, hingga evaluasi kinerja sistem, dijabarkan secara menyeluruh.

1.1. Preprocessing Data

Pada proses preprocessing dilakukan langkah-langkah case folding, tokenizing, stopword removal dan stemmig menggunakan pyhton untuk meningkatkan kualitas data input kedalam model seperti berikut ini:

```
# Inisialisasi stopwords dan stemmer
stop_words = set(stopwords.words('indonesian'))
factory = StemmerFactory()
stemmer = factory.create_stemmer()

# Fungsi preprocessing
def preprocess(text):
    text = text.lower()  # ubah ke huruf kecil
    text = text.translate(str.maketrans('', '', string.punctuation))  # hapus tanda baca
    tokens = text.split()  # tokenisasi
    tokens = [word for word in tokens if word not in stop_words]  # hapus stopword
    tokens = [stemmer.stem(word) for word in tokens]  # stemming
    return ' '.join(tokens)

# Baca ulang file tanpa header
df = pd.read_csv('skripsi.csv', encoding='latin1', header=None)

# Lakukan pemisahan isi berdasarkan koma
df = df[0].str.split(',', expand=True)
```

	Judul	judul_clean
0	SISTEM INFORMASI MANAJEMEN RUSUNAWA POLITEKNIK...	sistem informasi manajemen rusunawa politeknik...
1	RANCANG BANGUN SISTEM INFORMASI TABUNGAN AMAL ...	rancang bangun sistem informasi tabung amal ma...
2	Penerapan Automatic location pada website pela...	terap automatic location website lapor rusa as...
3	Pengembangan Galeri Sastra Berbasis Multiplatf...	kembang galeri sastra bas multiplatform metode...
4	PENERAPAN METODE EXTREME PROGRAMMING PADA APLI...	terap metode extreme programming aplikasi cata...

Gambar 1. Praproses Data

1.2. Training Data

Sebelum melakukan training data, untuk memastikan bahwa distribusi kelas (label) pada data latih dan uji serupa dengan distribusi kelas pada seluruh dataset digunakan metode stratified sampling dataset dibagi menjadi data latih dan data uji dengan perbandingan 80:20, berikut ini:

```
from sklearn.model_selection import train_test_split

# Bagi data menjadi latih dan uji dengan stratified sampling
X_train, X_test, y_train, y_test = train_test_split(df['judul_clean'], df['Konsentrasi'],
test_size=0.2, random_state=42, stratify=df['Konsentrasi'])
```

Gambar 2. Data Latih dan Uji

1.3. Mengubah teks menjadi vektor numerik menggunakan TF-IDF.

Berikut ini mengubah teks judul skripsi menjadi vektor fitur numerik menggunakan TF-IDF:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report

# Ubah teks menjadi fitur numerik menggunakan TfidfVectorizer
vectorizer = TfidfVectorizer(max_features=5000)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)
```

Gambar 3. Konversi Teks ke Vektor Numerik

Hasilnya seperti berikut ini:

	Kata	TF-IDF
0	anti	0.309648
1	based	0.309648
2	fake	0.309648
3	gps	0.309648
4	presensi	0.309648
5	sekolah	0.309648
6	location	0.282942
7	pegawai	0.282942
8	service	0.282942
9	dasar	0.249297

1.4. Evaluasi Model menggunakan Naïve Bayes
Algoritma Naïve Bayes diterapkan untuk melatih model klasifikasi dan mengukur akurasinya, yang selanjutnya digunakan dalam proses klasifikasi judul skripsi baru, sebagai berikut:

```
# Bangun model Naive Bayes
model = MultinomialNB()
model.fit(X_train_tfidf, y_train)

# Lakukan prediksi dengan data uji
y_pred = model.predict(X_test_tfidf)

# Evaluasi model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy * 100:.2f}%')

# Tampilkan laporan klasifikasi
print(classification_report(y_test, y_pred))
```

Hasil dari akurasi, Precision, Recall, F1-Scorenya sebagai berikut:

```
Accuracy: 65.00%
                                precision    recall  f1-score   support

          Full Stack (Web) Developer       0.50      0.80      0.62         5
                    Mobile Developer       0.00      0.00      0.00         2
                    Software Engineer       0.73      1.00      0.84         8
Software Quality Assurance Engineer       1.00      0.50      0.67         2
                         Technopreneur       0.00      0.00      0.00         2
                         UI/UX Designer       0.00      0.00      0.00         1

                               accuracy                           0.65        20
                              macro avg       0.37      0.38      0.35        20
                           weighted avg       0.52      0.65      0.56        20
```

Pembahasan

Hasil evaluasi model klasifikasi menunjukkan akurasi 65%, yang artinya dari 20 judul skripsi yang diuji, 13 judul berhasil diklasifikasikan dengan benar. Berikut ini hasil Prescision, Recall dan F1-Score Perkonsentrasinya:

Table 1. Data Akurasi

Konsentrasi	Precision	Recall	F1-score	Support
Full Stack (Web) Developer	0.50	0.80	0.62	5
Mobile Developer	0.00	0.00	0.00	2
Software Engineer	0.73	1.00	0.84	8
Software Quality Assurance	1.00	0.50	0.67	2
Technopreneur	0.00	0.00	0.00	2
UI/UX Designer	0.00	0.00	0.00	1

Berdasarkan tabel di atas konsentrasi Software Engineer memiliki performa terbaik dengan recall 1.00 yang artinya semua judul berhasil terprediksi dengan benar dan precision 0.73 yang menunjukkan rata-rata prediksi untuk kategori ini juga benar. Konsentrasi Full Stack Developer memiliki hasil Precision 0.50 artinya dari semua yang diprediksi sebagai “Full Stack”, hanya setengahnya yang benar, tapi recall 0.80 menunjukkan sebagian besar data aslinya berhasil dikenali. Untuk konsentrasi Software Quality Assurance menunjukkan precision 1.00 yaitu semua prediksi benar, tapi recall 0.50 berarti hanya setengah dari data yang seharusnya SQA berhasil ditemukan. Selanjutnya untuk konsentrasi Mobile Developer, Technopreneur, dan UI/UX Designer memiliki precision dan recall 0.00. Artinya tidak ada satu pun prediksi yang benar untuk kategori ini.

4. Kesimpulan

Penelitian ini menunjukkan bahwa algoritma Naive Bayes dapat digunakan untuk mengklasifikasikan judul skripsi ke dalam enam konsentrasi program studi dengan akurasi sebesar 65%. Konsentrasi Software Engineer memiliki hasil terbaik, sementara beberapa konsentrasi lain belum terklasifikasi dengan baik karena jumlah data yang terbatas dan variasi kata yang tinggi.

Teknik praproses seperti tokenizing, stopword removal, dan TF-IDF terbukti membantu

meningkatkan kualitas data. Namun, untuk hasil yang lebih akurat dan merata, dibutuhkan penambahan data serta eksplorasi algoritma lain seperti Decision Tree atau Neural Network.

1. Naive Bayes cukup efektif untuk klasifikasi judul skripsi, terutama jika data tiap konsentrasi seimbang.
2. Praproses data sangat membantu, tapi pengembangan model dan algoritma lain masih diperlukan untuk hasil yang lebih optimal.

5. Daftar Pustaka

Dhuhita, W. M. P., Darmawan, M. F. K. A., Triana, L., & Ankisqiantari, N. (2022). Perbandingan Algoritma Supervised Learning untuk Klasifikasi Judul Skripsi Berdasarkan Bidang Dosen. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(2), 427–437. https://doi.org/10.28932/jutisi.v8i2.4960

Ilmiah, J., & Komputer, I. (2024). *Penerapan Algoritma Naive Bayes Classifier*. *3*(1), 15–22.

Kurniadi, D. (2024). The application of naive bayes method for final project topic selection within the project-based learning framework in the data mining course. *Jurnal EDUCATIO: Jurnal Pendidikan Indonesia*, *10*(1), 243. https://doi.org/10.29210/1202423794

Mubarak, R., Hanafi, M., & Sasongko, D. (2024). Komparasi Performa Naive Bayes Gaussian dan K-NN Untuk Prediksi Kelulusan Mahasiswa dengan CRISP-DM. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(6), 2982–2991. https://doi.org/10.30865/klik.v4i6.1924

Nuraeni, R., Sudiarjo, A., & Rizal, R. (2021). Perbandingan Algoritma Naïve Bayes Classifier dan Algoritma Decision Tree untuk Analisa Sistem Klasifikasi Judul Skripsi. *Innovation in Research of Informatics (INNOVATICS)*, *3*(1), 26–31. https://doi.org/10.37058/innovatics.v3i1.2976

Sukriadi, S., Ismail, I., & Andzar, A. M. (2023). Penerapan Text Mining Dalam Klasifikasi Judul Skripsi Yang Diusulkan Mahasiswa Menggunakan Metode Naïve Bayes. *Jurnal Ilmiah Sistem Informasi Dan Teknik Informatika (JISTI)*, *6*(2), 184–196. https://doi.org/10.57093/jisti.v6i2.174

Suppa, R. (2023). Comparative Performance Evaluation Results of Classification Algorithm in Data Mining to Identify Types of Glass Based on Refractive Index and It's Elements. *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, *8*(1), 20. https://doi.org/10.51557/pt_jiit.v8i1.1705

Thesis, B. (2012). *Erik Lux Feature selection for text classification with Naive Bayes*.