Klasifikasi Judul Proyek Aplikasi Konsentrasi Menggunakan Algoritma Naive Bayes

Ardimansyah[1], Mudarsep[2], Baharuddin Rahman[3], Muhammad Khaddafi Tayeb[4]
Universitas Dipa Makassar
Jl. Perintis Kemerdekaan Km.9; Telp. 0411- 587194
e-mail: [1]ardiman@undipa.ac.id,[2]mudarsep@undipa.ac.id,[3]baharuddin.rahman@undipa.ac.id,
[4]khaddafy_thayyeb@undipa.ac.id

Abstrak

Klasifikasi judul proyek aplikasi berdasarkan bidang konsentrasi menjadi tantangan bagi institusi pendidikan dan perusahaan teknologi karena variasi kata dan struktur judul yang beragam. Proses klasifikasi manual memerlukan waktu lama dan rentan terhadap kesalahan subjektif. Penelitian ini mengusulkan penggunaan algoritma Naïve Bayes dengan teknik pemrosesan bahasa alami (NLP) untuk mengotomatiskan klasifikasi berdasarkan pola dalam data historis. Hasil penelitian menunjukkan bahwa model memiliki akurasi sebesar 42.42%. Beberapa kategori, seperti Pemerintahan & Hukum serta Pariwisata & Hospitality, menunjukkan precision dan recall yang tinggi, sedangkan kategori lain kurang terdeteksi akibat keterbatasan data. Nilai macro average precision sebesar 0.33 dan recall 0.25 mengindikasikan model masih kurang optimal dalam mengenali seluruh kategori. Untuk meningkatkan akurasi, diperlukan peningkatan jumlah data, teknik balancing, serta eksplorasi model lain seperti SVM atau Random Forest. Dengan implementasi sistem klasifikasi otomatis ini, pengelolaan data proyek dapat dilakukan lebih efisien, serta mendukung analisis tren penelitian di lingkungan akademik maupun industri.

Kata kunci: Klasifikasi judul proyek, Naïve Bayes, NLP, machine learning, analisis teks.

Abstract

Classifying application project titles based on their concentration field is a challenge for educational institutions and technology companies due to variations in wording and title structures. Manual classification is time-consuming and prone to subjective errors. This study proposes the use of the Naïve Bayes algorithm with Natural Language Processing (NLP) techniques to automate classification based on patterns in historical data. The results show that the model achieves an accuracy of 42.42%. Certain categories, such as Government & Law and Tourism & Hospitality, exhibit high precision and recall, while others are poorly detected due to limited data. The macro average precision of 0.33 and recall of 0.25 indicate that the model is still suboptimal in recognizing all categories. To improve accuracy, data augmentation, category balancing, and exploration of more advanced models such as Support Vector Machine (SVM) or Random Forest are recommended. By implementing this automated classification system, project data management can be more efficient, supporting trend analysis in both academic and industrial environments.

Keywords: *Project title classification, Naïve Bayes, NLP, machine learning, text analysis.*

1. Pendahuluan

Dalam dunia akademik dan penelitian, banyak institusi pendidikan dan perusahaan teknologi menghadapi tantangan dalam mengelompokkan proyek atau aplikasi berdasarkan bidang konsentrasinya. Judul proyek aplikasi sering kali memiliki variasi kata dan struktur yang berbeda, sehingga sulit untuk mengklasifikasikan secara otomatis. Proses klasifikasi manual membutuhkan waktu yang lama dan rentan terhadap kesalahan subjektif. Oleh karena itu, diperlukan suatu metode yang dapat mengotomatiskan proses klasifikasi judul proyek aplikasi dengan tingkat akurasi yang tinggi.

Permasalahan utama dalam klasifikasi judul proyek aplikasi adalah perbedaan terminologi yang digunakan dalam penulisan judul. Satu proyek yang berfokus pada kecerdasan buatan, misalnya, bisa memiliki judul yang berbeda-beda tergantung pada aspek teknologi yang ditekankan. Selain itu, metode

klasifikasi tradisional yang berbasis aturan sering kali tidak fleksibel dan kurang efektif dalam menangani berbagai variasi bahasa yang digunakan dalam judul proyek. Tanpa sistem klasifikasi yang baik, proses pencarian, analisis, dan manajemen proyek menjadi kurang efisien.

Untuk mengatasi permasalahan tersebut, penelitian ini mengusulkan penggunaan algoritma Naïve Bayes dalam klasifikasi judul proyek aplikasi berdasarkan konsentrasinya. Algoritma Naïve Bayes merupakan metode pembelajaran mesin berbasis probabilistik yang telah terbukti efektif dalam tugas klasifikasi teks. Dengan menerapkan teknik pemrosesan bahasa alami (NLP) untuk ekstraksi fitur dari judul proyek, algoritma ini dapat menganalisis pola dalam data historis dan memprediksi kategori yang paling sesuai untuk setiap judul proyek baru.

Dengan diterapkannya sistem klasifikasi otomatis menggunakan algoritma Naïve Bayes, diharapkan proses identifikasi dan pengelompokan proyek berdasarkan bidang konsentrasi dapat dilakukan dengan lebih cepat, akurat, dan efisien. Sistem ini juga dapat membantu institusi pendidikan dalam mengelola data proyek mahasiswa serta mendukung pengambilan keputusan dalam analisis tren penelitian. Selain itu, hasil klasifikasi yang lebih akurat akan meningkatkan aksesibilitas dan pencarian proyek di berbagai platform akademik maupun industri.

2. Metode Penelitian

Penelitian ini menggunakan metode klasifikasi teks dengan algoritma Naïve Bayes untuk mengelompokkan judul-judul proyek ke dalam kategori tertentu. Proses penelitian terdiri dari beberapa tahap utama, yaitu pengumpulan data, preprocessing teks, pembagian data, ekstraksi fitur, pelatihan model, evaluasi model, dan interpretasi hasil.

2.1 Pemilihan Dataset

Dataset adalah kumpulan data terstruktur dan terorganisir yang digunakan untuk analisis dan pemodelan. Pemilihan dataset yang tepat sangat penting untuk menghasilkan model yang akurat dan berkinerja baik[1].Data yang digunakan dalam penelitian ini berupa kumpulan judul proyek yang diperoleh dari berbagai sumber. Data disimpan dalam format Excel (.xlsx), yang kemudian diolah menggunakan bahasa pemrograman Python pada platform Google Colaboratory. Data ini terdiri dari dua kolom utama:
Judul: Berisi teks judul proyek yang akan diklasifikasikan.
Kategori: Label yang menunjukkan kategori dari masing-masing judul.

2.2 Preprocessing Data

Preprocessing menjadi tahap awal dalam klasifikasi teks untuk mempersiapkan data teks sebelum digunakan pada proses lainnya [2]. Tahap ini bertujuan untuk membersihkan teks agar lebih optimal dalam proses klasifikasi. Langkah-langkah preprocessing meliputi: Case Folding yaitu mengubah semua huruf menjadi huruf kecil. Tokenization yaitu memisahkan teks menjadi kata-kata individu. Stopword Removal yaitu menghapus kata-kata umum yang tidak memiliki makna signifikan dalam klasifikasi. Stemming yaitu mengubah kata menjadi bentuk dasarnya menggunakan algoritma Porter Stemmer. Handling Missing Values yaitu menghapus atau mengganti data yang kosong (NaN).

2.3 Split Data

Split data adalah pembagian data training dan testing yaitu dengan pembagian 80% untuk training dan 20% untuk testing [3].Setelah preprocessing, data dibagi menjadi data latih (train data) dan data uji (test data) dengan perbandingan 80:20. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk mengukur performa model.

2.4 Ekstraksi Fitur

Ekstraksi fitur merupakan proses mengubah data yang berupa teks yang tidak terstruktur menjadi data yang terstruktur untuk dapat diproses selanjutnya ke tahapan klasifikasi[4].Untuk mengubah teks menjadi representasi numerik yang dapat digunakan oleh algoritma Naïve Bayes, digunakan teknik TF-IDF (Term Frequency - Inverse Document Frequency) yang menghitung bobot setiap kata dalam dokumen berdasarkan frekuensi kemunculannya.

2.5 Pelatihan Model Naive Bayes

Naive Bayes merupakan pengklasifikasi probabilitas yang didasarkan pada Teorema Bayes. Keunikan dari Naive Bayes terletak pada asumsi bahwa setiap atribut bersifat bebas (independent). Algoritma ini dapat dilatih secara efisien dalam pembelajaran terawasi dan memiliki keuntungan karena membutuhkan sejumlah kecil data pelatihan untuk memperkirakan parameter yang diperlukan untuk klasifikasi[5].Pada tahap ini, model Multinomial Naive Bayes diterapkan untuk mempelajari pola dari data latih. Model ini dipilih karena efektif dalam klasifikasi teks, khususnya dalam mengatasi masalah dengan fitur berbasis probabilitas

2.6 Evaluasi Model

Setelah model dilatih, dilakukan evaluasi menggunakan metrik berikut:

1. Akurasi: Persentase prediksi yang benar dari total data uji.
2. Precision, Recall, dan F1-Score: Mengukur kinerja model dalam klasifikasi dengan lebih rinci.

2.6 Interpretasi Hasil

Hasil klasifikasi dianalisis untuk mengetahui sejauh mana model mampu mengelompokkan judul proyek dengan benar. Jika akurasi masih rendah, dilakukan optimasi dengan tuning parameter atau penyesuaian preprocessing data.

3. Hasil dan Pembahasan

Dalam penelitian ini akan digunakan data sebanyak 492 judul proyek aplikasi yang akan akan dikelompokkan menggunakan algoritma Naïve Bayes.

Berikut adalah contoh lengkap hasil clustering dalam bentuk tabel untuk 20 data judul:

Tabel 1Data Judul Proyek Aplikasi

No.	Judul	Kategori
1	Aplikasi Perhitungan Kalender Kehamilan	Kesehatan
2	Rekomendasi Tempat Kulineran dan Cafe di Lokasi	Sistem Informasi & Administrasi
3	Aplikasi Pengenalan Objek Wisata Sulawesi Selatan Berbasis WEB	Pariwisata & Hospitality
4	Pembuatan Website Quizme	Teknologi Pendidikan
5	Aplikasi Call & Cuts	Layanan Digital
6	Aplikasi Pusat Pencarian Pekerjaan Online	Sistem Informasi & Administrasi
7	Manajemen Kelas Musik	Teknologi Pendidikan
8	Aplikasi Pemanggil Tukang Cuci Kendaraan Kerumah (WEB)	Layanan Digital
9	Pengembangan Chat Bot WhatsApp untuk Pencarian Produk	Teknologi & Kecerdasan Buatan
10	Aplikasi Sistem Informasi Pendaftaran Satgas Pariwisata Kota Luwu Utara berbasis web	Pariwisata & Hospitality
11	Sistem Informasi Penjadwalan Guru Berbasis Web	Teknologi Pendidikan
12	Aplikasi Netflix Movie Berbasis Mobile Android	Media
13	Sistem Informasi Pencarian Bengkel dengan Cepat dan Mudah Berbasis Web	Layanan Digital
14	Sistem Pemantauan dan Pengendali Jarak Jauh Berbasis Website dan IoT secara Real Time yang Dilengkapi Keamanan Sensor Sidik Jari	IoT & Automasi
15	Monitoring Ketinggian Angkutan Barang pada Mobil Pick Up untuk Menghindari Over Dimension Over Loading (ODOL) pada Jembatan Timbang Maccopa Maros	Transportasi & Logistik
…	…	…
488	Sistem Informasi Puskesmas Berbasis Web	Kesehatan

No.	Judul	Kategori
489	Sistem Informasi Pembuatan Kartu Pencari Kerja (Kartu Kuning) Pada Dinas Sosial Tenaga Kerja Dan Transmigrasi	Sistem Informasi & Administrasi
490	Layanan home service berbasis web menggunakan framework Laravel	Jasa
491	Sistem informasi pemasaran makanan berbasis web	Kuliner & Reservasi

Selanjutnya berdasarkan data yang ada maka akan dilakukan Langkah-langkah berikut:

1. Preprocesing

Melakukan proprocesing data untuk membersihkan data yang akan diekstraksi.

Tabel 2 Data Setelah Preprocessing 20 data awal

	Judul	Judul Clean
0	Aplikasi Perhitungan Kalender Kehamilan	aplikasi perhitungan kalender kehamilan
1	Rekomendasi Tempat Kulineran dan Cafe di Lokasi	rekomendasi kulineran cafe lokasi
2	Aplikasi Pengenalan Objek Wisata Sulawesi Sela...	aplikasi pengenalan objek wisata sulawesi sela...
3	Pembuatan Website Quizme	pembuatan website quizme
4	Aplikasi Call & Cuts	aplikasi call cuts
5	Aplikasi Pusat Pencarian Pekerjaan Online	aplikasi pusat pencarian pekerjaan online
6	Manajemen Kelas Musik	manajemen kelas musik
7	Aplikasi Pemanggil Tukang Cuci Kendaraan Kerum...	aplikasi pemanggil tukang cuci kendaraan kerum...
8	Pengembangan Chat Bot WhatsApp untuk Pencarian...	pengembangan chat bot whatsapp pencarian produk
9	Aplikasi Sistem Informasi Pendaftaran Satgas P...	aplikasi sistem informasi pendaftaran satgas p...
10	Sistem Informasi Penjadwalan Guru Berbasis Web	sistem informasi penjadwalan guru berbasis web
11	Aplikasi Netflix Movie Berbasis Mobile Android	aplikasi netflix movie berbasis mobile android
12	Sistem Informasi Pencarian Bengkel dengan Cepa...	sistem informasi pencarian bengkel cepat mudah...
13	Sistem Pemantauan dan Pengendali Jarak Jauh Be...	sistem pemantauan pengendali jarak berbasis we...
14	Monitoring Ketinggian Angkutan Barang pada Mob...	monitoring ketinggian angkutan barang mobil pi...
487	Sistem Informasi Puskesmas Berbasis Web	sistem informasi puskesmas berbasis web
488	SISTEM INFORMASI PEMBUATAN KARTU PENCARI KERJA...	sistem informasi pembuatan kartu pencari kerja...
489	Layanan home service berbasis web menggunakan ...	layanan home service berbasis web framework la...
490	Sistem informasi pemasaran makanan berbasis web	sistem informasi pemasaran makanan berbasis web

2. Hasil Ekstraksi

Melakukan Ekstraksi data pada data yang telah dipreprocesing:

Tabel 3TF-IDF Hasil Ekstraksi (5 data pertama)

11	21	absensi	acak	...	warga	webinar	website	wedding	whatsapp	wisata
0	0	0	0	...	0	0	0	0	0	0
1	0	0	0	...	0	0	0	0	0	0
2	0	0	0	...	0.129468	0	0	0	0	0.353056
3	0	0	0	...	0	0	0.403722	0	0	0
4	0	0	0	...	0	0	0	0	0	0

3. Split Data

Berikut adalah Split data atau pembagian data yang akan digunakan:

Tabel 4Split Data

	Jenis Data	Jumlah
0	Data Training	392
1	Data Testing	99

4. Hasil Prediksi

Melakukan pengujian prediksi data yang dilatih.

Tabel 5Hasil Prediksi pada Data Uji

	Asli	Prediksi
452	Sumber Daya Manusia & Organisasi	Edukasi & Akademik
84	Layanan Digital	Sistem Informasi & Administrasi
434	Peminjaman & Penyewaan	Transportasi & Logistik
474	Properti & Penyewaan	Edukasi & Akademik
428	Produktivitas & Manajemen	Produktivitas & Manajemen
312	Edukasi & Akademik	E-Commerce & Retail
30	Produktivitas & Manajemen	E-Commerce & Retail
220	Kesehatan	Kesehatan
484	Transportasi & Logistik	Transportasi & Logistik
231	Transportasi & Logistik	Transportasi & Logistik
9	Pariwisata & Hospitality	Edukasi & Akademik
124	Transportasi & Logistik	E-Commerce & Retail
422	Produktivitas & Manajemen	Edukasi & Akademik
204	Pemerintahan & Hukum	Pemerintahan & Hukum
360	Teknologi & Kecerdasan Buatan	E-Commerce & Retail
70	Media	Edukasi & Akademik
430	Edukasi & Akademik	Edukasi & Akademik
431	Pariwisata & Hospitality	Pariwisata & Hospitality
364	Lingkungan & Smart City	Edukasi & Akademik
211	Teknologi Pendidikan	Teknologi Pendidikan

5. Evaluasi Model

Melakukan evaluasiterhadap model.

Tabel 6Akurasi Model

	Metrik	Nilai
0	Akurasi Model	0.424242

Nilai akurasi 0.424 (42.42%) menunjukkan bahwa model Naïve Bayes berhasil mengklasifikasikan sekitar 42.42% dari total data uji dengan benar.

Tabel 7 Laporan Klasifikasi

	precision	recall	f1-score	support
E-Commerce & Retail	0.096774	0.6	0.166667	5
Edukasi & Akademik	0.428571	0.75	0.545455	12
IoT & Automasi	0	0	0	1
Jasa	0	0	0	1
Karier & Lowongan Kerja	0	0	0	1
Kesehatan	1	0.428571	0.6	14
Keuangan & Fintech	0.75	0.428571	0.545455	7
Kuliner & Reservasi	1	0.5	0.666667	2
Layanan Digital	0	0	0	3
Lingkungan & Smart City	0	0	0	1
Media	0	0	0	3
Olahraga & Rekreasi	0	0	0	3
Otomotif & Layanan Jasa	0	0	0	1
Pariwisata & Hospitality	0.625	0.833333	0.714286	6
Pemerintahan & Hukum	1	0.75	0.857143	4
Peminjaman & Penyewaan	0	0	0	2
Pendidikan	0	0	0	1
Produktivitas & Manajemen	0.75	0.5	0.6	6
Properti & Penyewaan	0	0	0	3
Sistem Informasi & Administrasi	0.5	0.4	0.444444	5

Sumber Daya Manusia & Organisasi	1	0.25	0.4	4
Survei & Analisis Data	0	0	0	1
Teknologi & Kecerdasan Buatan	0	0	0	1
Teknologi Pendidikan	1	0.25	0.4	4
Telekomunikasi	0	0	0	1
Transportasi & Logistik	0.333333	0.714286	0.454545	7
accuracy	0.424242	0.424242	0.424242	0.424242
macro avg	0.326295	0.246337	0.245948	99
weighted avg	0.524849	0.424242	0.412612	99

Hasil klasifikasi menggunakan Naïve Bayes menunjukkan bahwa model memiliki akurasi sebesar 42.42%, yang berarti hanya sekitar 42% data uji yang berhasil diprediksi dengan benar. Berdasarkan precision, recall, dan F1-score pada masing-masing kategori, terlihat bahwa beberapa kategori memiliki kinerja yang cukup baik, seperti Pemerintahan & Hukum dengan precision 1.00, recall 0.75, dan F1-score 0.86, serta Pariwisata & Hospitality dengan precision 0.625, recall 0.83, dan F1-score 0.71. Namun, ada banyak kategori yang memiliki nilai precision dan recall 0.00, seperti IoT & Automasi, Jasa, Layanan Digital, Media, dan Telekomunikasi. Hal ini menunjukkan bahwa model tidak mampu mengenali kategori-kategori tersebut, kemungkinan karena jumlah data yang terlalu sedikit atau pola yang sulit dipelajari oleh Naïve Bayes.

Selain itu, nilai macro average (rata-rata tanpa mempertimbangkan jumlah sampel) menunjukkan precision 0.33, recall 0.25, dan F1-score 0.25, yang mengindikasikan bahwa model masih kurang baik dalam mengenali pola secara umum di semua kategori. Sementara itu, weighted average (rata-rata dengan mempertimbangkan jumlah sampel per kategori) memiliki precision 0.52, recall 0.42, dan F1-score 0.41, yang menunjukkan bahwa model lebih cenderung memprediksi kategori dengan jumlah data lebih banyak.

4. Kesimpulan

Berdasarkan hasil penelitian ini, ada beberapa perbaikan yang dapat dilakukan untuk meningkatkan performa model. Salah satunya adalah menambah jumlah data pada kategori yang memiliki precision dan recall nol agar model dapat belajar pola dengan lebih baik. Selain itu, penggunaan model yang lebih kompleks seperti Support Vector Machine (SVM) atau Random Forest dapat dipertimbangkan untuk meningkatkan akurasi. Teknik balancing data, seperti oversampling kategori minoritas atau undersampling kategori mayoritas, juga dapat membantu model mengenali pola dengan lebih baik. Selain itu, tuning parameter pada algoritma Naïve Bayes juga bisa dilakukan untuk meningkatkan performa model. Dengan perbaikan-perbaikan ini, diharapkan model dapat memberikan hasil prediksi yang lebih akurat dan mampu mengenali lebih banyak kategori dengan baik.

Daftar Pustaka

[1] Rahayu, P. W., Sudipa, I. G. I., Suryani, S., Surachman, A., Ridwan, A., Darmawiguna, I. G. M.,& Maysanjaya, I. M. D. (2024). Buku Ajar Data Mining. PT. Sonpedia Publishing Indonesia.

[2] Khairunnisa, S., Adiwijaya, A., & Al Faraby, S. (2021). Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19). J. Media Inform. Budidarma, 5(2), 406.

[3] Astuti, Y., Wulandari, I. R., Putra, A. R., & Kharomadhona, N. (2022). Naïve Bayes untuk Prediksi Tingkat Pemahaman Kuliah Online Terhadap Mata Kuliah Algoritma Struktur Data. J. Edukasi Dan Penelit. Inform, 8(1), 28.
[4] Sari, S. N., Faisal, M. R., Kartini, D., Budiman, I., Saragih, T. H., & Muliadi, M. (2023). Perbandingan Ekstraksi Fitur dengan Pembobotan Supervised dan Unsupervised pada Algoritma Random Forest untuk Pemantauan Laporan Penderita COVID-19 di Twitter. Jurnal Komputasi, 11(1), 33-42.
[5] Maulana, M. I., Rahayudi, B., & Setiawan, N. Y. (2024). Analisis Pengelompokan Ulasan Pengguna menggunakan K-Means Clustering untuk Evaluasi Aplikasi My SAPK BKN. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 8(14).