Jurnal Teknik Informatika dan Teknologi Informasi
Volume 5, Nomor 2, Agustus 2025
E-ISSN: 2827-9387, P- ISSN: 2827-9379, Hal. 521-533
DOI: https://doi.org/10.55606/jutiti.v5i2.5693
Available Online at: https://journalshub.org/index.php/jutiti

Klasifikasi Dokumen Publik Berbasis NLP: Otomatisasi Proses Informasi Menuju Keterbukaan Data yang Adaptif dan Transparan

Retnowati Retnowati[1*], Veronica Lusiana[2], Eko Nur Wahyudi [3]
[1,2,3] Universitas Stikubank, Indonesia
Email : retnowati@edu.unisbank.ac.id [1]

Alamat: Jalan Tri Lomba Juang No. 1 Mugas Semarang
Korespondensi penulis: retnowati@edu.unisbank.ac.id *

Abstract. *In the era of public information disclosure, digital documents have become strategic assets in supporting transparent, accountable, and participatory governance. Effective management of these documents is essential to ensure that public information services are responsive and accessible. However, document classification tasks carried out by Public Information and Documentation Officers (PPID) still rely heavily on manual processes, which are time-consuming, inefficient, and prone to human error. To address this challenge, this study aims to develop an intelligent classification model for public documents using Artificial Intelligence (AI) and Natural Language Processing (NLP), integrated within the Data Lifecycle Management (DLM) framework. The proposed solution was designed using the Design Science Research (DSR) methodology and implemented through Agile development practices. Evaluation was conducted in a simulated laboratory environment that mirrors real-world PPID operations.The developed model leverages transformer-based architectures, particularly BERT (Bidirectional Encoder Representations from Transformers), and is compared against traditional algorithms such as Naive Bayes and K-Nearest Neighbors (KNN). Experimental results show that the BERT model achieves superior performance, with an accuracy of 89%, precision of 0.88, recall of 0.89, and F1-score of 0.88. These metrics confirm that Transformer-based models are highly effective for classifying public documents into categories of information accessibility: available at all times, periodic, immediate, and exempted from disclosure.This research highlights the potential of AI-powered classification to streamline public information services, reduce workload, and enhance compliance with information disclosure laws. The findings support national development priorities such as RPJMN 2025 by contributing to digital transformation in the public sector. The study also provides a replicable framework for other government agencies aiming to implement adaptive and transparent document classification systems.*

Keywords: *Document Classification; NLP; Naive Bayes; Public Information Disclosure; KNN*

Abstrak. Di era keterbukaan informasi publik, dokumen digital menjadi aset strategis dalam mendukung tata kelola pemerintahan yang transparan, akuntabel, dan partisipatif. Pengelolaan dokumen yang efektif sangat penting untuk memastikan layanan informasi publik yang responsif dan mudah diakses. Namun, proses klasifikasi dokumen yang dilakukan oleh Pejabat Pengelola Informasi dan Dokumentasi (PPID) masih mengandalkan cara manual yang memakan waktu, tidak efisien, dan rentan terhadap kesalahan manusia. Penelitian ini bertujuan untuk mengembangkan model klasifikasi dokumen publik berbasis Artificial Intelligence (AI) dan Natural Language Processing (NLP) yang terintegrasi dalam kerangka kerja Data Lifecycle Management (DLM). Pengembangan model dilakukan dengan pendekatan Design Science Research (DSR) dan metode Agile, serta diuji dalam lingkungan laboratorium simulasi yang mereplikasi operasi nyata PPID. Model yang dikembangkan memanfaatkan arsitektur berbasis Transformer, khususnya BERT (Bidirectional Encoder Representations from Transformers), dan dibandingkan dengan algoritma tradisional seperti Naive Bayes dan K-Nearest Neighbors (KNN). Hasil pengujian menunjukkan bahwa model BERT memiliki kinerja terbaik dengan akurasi sebesar 89%, precision 0,88, recall 0,89, dan F1-score 0,88. Temuan ini membuktikan bahwa teknologi berbasis Transformer sangat efektif dalam mengklasifikasikan dokumen ke dalam kategori aksesibilitas informasi publik: setiap saat, berkala, serta-merta, dan dikecualikan. Penelitian ini menunjukkan potensi kuat penggunaan AI dalam meningkatkan efisiensi layanan informasi publik, mengurangi beban kerja manual, serta mendukung pelaksanaan kebijakan nasional seperti RPJMN 2025. Selain itu, studi ini memberikan kontribusi dalam penerapan teknologi cerdas di sektor publik dan dapat dijadikan acuan dalam pengembangan sistem klasifikasi informasi yang adaptif, transparan, dan dapat direplikasi.

Kata kunci: Klasifikasi Dokumen; NLP, Naive Bayes; Keterbukaan Informasi Publik; KNN

Received: Juli 09, 2025; Revised: Juli 23, 2025; Accepted: Agustus 06, 2025; Published: Agustus 08, 2025

1. LATAR BELAKANG

Dalam era transformasi digital, data telah menjadi aset strategis untuk mewujudkan efisiensi, transparansi, dan akuntabilitas dalam tata kelola informasi publik. Komitmen terhadap pemerintahan terbuka ditegaskan melalui Undang-Undang No. 14 Tahun 2008 tentang Keterbukaan Informasi Publik, yang mewajibkan badan publik untuk menyediakan informasi secara cepat, tepat waktu, dan akurat. Pejabat Pengelola Informasi dan Dokumentasi (PPID) memiliki peran sentral dalam pengelolaan dokumen publik yang diklasifikasikan ke dalam kategori aksesibilitas: **tertutup (rahasia)** dan **terbuka** (Setiap Saat, Serta Merta, Berkala).

Namun, praktik di lapangan menunjukkan bahwa sebagian besar Organisasi Perangkat Daerah (OPD) masih menggunakan metode klasifikasi dokumen secara manual dan tidak terstruktur. Hal ini menghambat efektivitas manajemen daur hidup data (MDHD) dan seringkali menyebabkan ketidakefisienan, kesalahan klasifikasi, serta keterlambatan dalam pengambilan keputusan. Seiring meningkatnya volume data dan kompleksitas regulasi, dibutuhkan sistem cerdas dan adaptif untuk mendukung proses klasifikasi dokumen secara otomatis.

Teknologi Artificial Intelligence (AI) dan Natural Language Processing (NLP) menawarkan potensi besar untuk menjawab tantangan ini. Berbagai studi menunjukkan bahwa algoritma seperti K-Nearest Neighbors (KNN) dan Naive Bayes efektif dalam pengolahan citra dan analisis sentimen. Namun, penerapannya dalam klasifikasi dokumen publik berdasarkan kategori aksesibilitas masih belum dioptimalkan, terutama dalam bentuk model *end-to-end* yang sesuai dengan regulasi keterbukaan informasi di Indonesia.

Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model klasifikasi dokumen publik berbasis AI dan Natural Language Processing (NLP) yang mampu mengelompokkan dokumen ke dalam kategori aksesibilitas sesuai regulasi Undang-Undang Keterbukaan Informasi Publik. Model ini dikembangkan secara iteratif dengan pendekatan Design Science Research (DSR) dan metode Agile untuk memastikan proses pengembangan berjalan adaptif dan terukur. Evaluasi dilakukan melalui simulasi pengujian performa model pada data dokumen publik, sehingga diperoleh model klasifikasi yang akurat, efisien, dan siap untuk diintegrasikan ke dalam sistem layanan informasi publik.

Kebaruan dari penelitian ini terletak pada pengembangan model klasifikasi otomatis dokumen publik menggunakan pendekatan AI/NLP berbasis Transformer-BERT yang secara khusus disesuaikan dengan kategori aksesibilitas dokumen menurut UU No. 14 Tahun 2008. Berbeda dengan studi sebelumnya yang hanya menerapkan NLP untuk analisis sentimen atau

klasifikasi umum, penelitian ini menyajikan pendekatan *end-to-end* untuk klasifikasi dokumen publik dalam konteks regulasi keterbukaan informasi di Indonesia.

2. KAJIAN TEORITIS

Keterbukaan Informasi Publik dan Kategori Aksesibilitas

Keterbukaan informasi publik merupakan amanat undang-undang yang menegaskan bahwa setiap warga negara berhak untuk mengakses informasi dari badan publik secara cepat dan akurat. Informasi tersebut diklasifikasikan ke dalam empat kategori utama: Berkala, Setiap Saat, Serta Merta, dan Dikecualikan (PP No. 61 Tahun 2010 Pelaksanaan Undang-Undang Nomor 14 Tahun 2008 Tentang Keterbukaan Informasi Publik, 2010; UU KIP No. 14 Tahun 2008, 2008). Beberapa penelitian terdahulu telah menyoroti urgensi klasifikasi informasi ini dalam konteks tata kelola informasi publik (Retnowati et al., 2021; Retnowati Retnowati et al., 2022), termasuk pentingnya peran PPID dalam memastikan aksesibilitas data secara transparan dan akuntabel.

Namun, banyak OPD di tingkat daerah yang melakukan pengelolaan dokumen secara manual (Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Republik Indonesia, 2024; Retnowati et al., 2018; Retnowati Retnowati et al., 2022; Retnowati Retnowati, Listiyono, Anwar, et al., 2019; Retnowati Retnowati, Listiyono, Purwatiningtyas, et al., 2019), yang memiliki risiko terhadap ketidakefisienan, kesalahan klasifikasi dan sengketa. Kondisi ini menggarisbawahi kebutuhan akan sistem klasifikasi otomatis yang dapat menjawab tantangan akurasi dan volume data yang besar.

Manajemen Daur Hidup Data (MDHD) dalam Tata Kelola Dokumen

Manajemen Daur Hidup Data (MDHD) merupakan kerangka pengelolaan informasi mulai dari penciptaan, penyimpanan, pemanfaatan, hingga pemusnahan dokumen (Felix C Aguboshim et al., 2023; Linthorst & de Waal, 2020; Prasetyo A et al., 2019; Sternad Zabukovšek et al., 2023). Dalam konteks keterbukaan informasi publik, klasifikasi merupakan langkah awal kritis untuk menentukan perlakuan dokumen dalam siklus hidupnya (Bennich, 2024; Sofyan et al., 2024; Wardhani et al., 2023). Penelitian sebelumnya telah menghasilkan mengembangkan model manajemen informasi publik berbasis kerangka SSM (R. Retnowati et al., 2019), namun belum menerapkan otomatisasi klasifikasi dengan pendekatan kecerdasan buatan.

Kebutuhan pengelolaan data berbasis sistem menjadi semakin mendesak seiring meningkatnya jumlah dokumen digital yang masuk dalam sistem PPID. Oleh karena itu, diperlukan pendekatan yang mampu mengintegrasikan sistem klasifikasi ke dalam proses MDHD secara otomatis dan kontekstual.

Natural Language Processing (NLP) dalam Klasifikasi Teks

Natural Language Processing (NLP) merupakan bagian dari kecerdasan buatan yang berfokus pada pemrosesan bahasa alami. Dalam tugas klasifikasi dokumen, NLP digunakan untuk mengekstraksi fitur dari teks dan membentuk representasi semantik (Pichiyan et al., 2023). NLP sederhana telah diterapkan dalam penelitian sebelumnya untuk analisis sentimen dan kebutuhan masyarakat, menggunakan Naive Bayes (Aryasatya & Lusiana, 2024; Muslimin & Lusiana, 2023), namun belum diterapkan untuk konteks regulasi seperti UU KIP.

Kajian oleh (Khurana et al., 2023) menunjukkan bahwa NLP modern telah berkembang jauh melampaui tokenisasi dan stemming, dan kini memanfaatkan pembelajaran kontekstual untuk memahami makna yang tersembunyi dalam dokumen. Potensi ini dapat dimanfaatkan untuk memahami isi dokumen publik dan menentukan klasifikasinya secara lebih akurat.

Transformer-BERT sebagai Representasi Semantik Dokumen

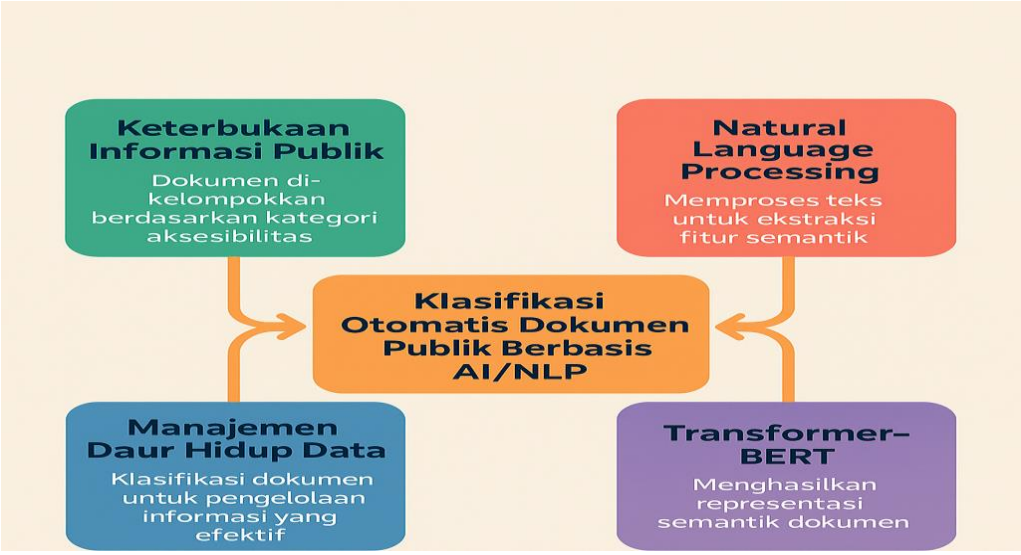
Transformer-BERT (Bidirectional Encoder Representations from Transformers) telah menjadi model unggulan dalam berbagai tugas NLP karena kemampuannya menangkap konteks kata secara bidirectional. Dalam klasifikasi dokumen, token [CLS] dari output BERT sering digunakan sebagai representasi dokumen secara keseluruhan. BERT dapat digunakan untuk mengekstraksi fitur teks dalam rangka membantu klasifikasi dokumen yang tidak terstruktur (Devlin et al., 2019; Pichiyan et al., 2023).

Namun, sejauh ini belum ada penelitian di Indonesia yang memanfaatkan representasi BERT untuk klasifikasi dokumen publik berdasarkan kategori aksesibilitas sesuai UU KIP. Hal ini menjadi celah yang sangat signifikan, mengingat regulasi nasional memerlukan pemahaman kontekstual atas isi dokumen.

Algoritma Klasifikasi: Naive Bayes dan K-Nearest Neighbors

Naive Bayes merupakan algoritma klasifikasi probabilistik yang telah terbukti efektif dalam berbagai tugas NLP, termasuk klasifikasi teks pendek (Liu, 2023). Sementara itu, K-Nearest Neighbors (KNN) menggunakan pendekatan kedekatan vektor untuk menentukan kategori dokumen (Dubey & R, 2024). Penelitian sebelumnya telah menggunakan kedua algoritma ini untuk pengenalan pola dalam citra (Lusiana, V., Al Amin, I. H., Hartono, B., & Kristianto, 2019), namun penerapannya untuk dokumen pemerintah masih sangat terbatas.

Kedua algoritma ini dipilih dalam penelitian ini karena kesederhanaannya, kemampuannya menangani data multikategori, serta kompatibilitasnya dengan representasi *embedding* dari BERT.



Gambar 2.1 Keterkaitan Teori Yang Melandasi Klasifikasi Otomatis Dokumen Publik

Dari penjelasan tentang kajian teori dapat dijelaskan keterkaitan teori-teori tersebut yang dipergunakan untuk menjawab masalah penelitian seperti tampak pada gambar 2.1. Penelitian ini mengembangkan model klasifikasi dokumen publik berbasis NLP yang secara eksplisit mengacu pada kategori aksesibilitas UU KIP. Penelitian ini juga mengintegrasikan Transformer-BERT sebagai ekstraktor fitur semantik dalam dokumen formal pemerintahan.

3. METODE PENELITIAN

Desain Penelitian

Jenis penelitian ini merupakan **penelitian rekayasa sistem informasi (information system design research)** yang bertujuan membangun model klasifikasi dokumen publik berbasis AI dan Natural Language Processing (NLP). Model dikembangkan secara end-to-end, mulai dari preprocessing dokumen, ekstraksi fitur semantik menggunakan BERT, hingga klasifikasi menggunakan algoritma supervised learning.

Populasi dan Sampel Penelitian

Populasi dalam penelitian ini adalah seluruh dokumen publik dari badan publik yang dikelola oleh PPID. Sampel dipilih secara **purposive** dari 8 badan publik di Provinsi Jawa Tengah yang mewakili kategori informatif, cukup informatif, dan kurang informatif menurut KIP Award. Setiap badan publik menyumbang sejumlah dokumen ($n \geq 30$) yang telah

diklasifikasikan secara manual oleh PPID, yang selanjutnya digunakan sebagai data latih dan uji.

Teknik dan Instrumen Pengumpulan Data

Data dikumpulkan melalui: studi dokumentasi terhadap dokumen publik aktual diperoleh secara tersebar di berbagai sumber situs PPID yang dinformasikan bago kepentingan masyarakat .

Tahapan dan Alat Analisis Data

a. Preprocessing Dokumen

Melibatkan tokenisasi, stopword removal, dan normalisasi teks. Tahapan ini menggunakan Python dan pustaka NLP seperti spaCy dan NLTK.

b. Ekstraksi Fitur dengan BERT

Setiap dokumen ddd diubah menjadi representasi vektor menggunakan model **BERT**. Representasi yang digunakan adalah vektor dari token **[CLS]**:

$$\vec{d} = \text{BERT}_{[\text{CLS}]}(d) \quad \text{Rumus (1)}$$

di mana $\vec{d} \in \mathbb{R}^n$ merepresentasikan dokumen dalam bentuk embedding berdimensi tinggi.

c. Klasifikasi Dokumen

Dokumen yang telah direpresentasikan sebagai vektor diklasifikasikan menggunakan dua algoritma:

1. **Naive Bayes**

   Algoritma ini menghitung probabilitas kemunculan dokumen dalam setiap kategori C_k berdasarkan fitur vektor.

$$P(C_k|\vec{d}) = \frac{P(C_k)\prod_{i=1}^{n} P(d_i|C_k)}{P(\vec{d})} \quad \text{Rumus (2)}$$

   Hasil akhir adalah:

$$\hat{C} = \arg\max_{C_k} P(C_k|\vec{d}) \quad \text{Rumus (3)}$$

2. **K-Nearest Neighbors (KNN)**

   Dokumen $\vec{d}$ dibandingkan dengan vektor dokumen lainnya menggunakan jarak Euclidean.

$$\text{distance}(\vec{d}, \vec{d}_j) = \sqrt{\sum_{i=1}^{n} (d_i - d_{j,i})^2} \quad \text{Rumus (4)}$$

Label dokumen diputuskan berdasarkan voting dari k tetangga terdekat.

d. Evaluasi Kinerja Model

Kinerja klasifikasi dievaluasi menggunakan metrik berdasarkan akurasi, presisi, recall, dan F1-Score.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Rumus (5)}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Rumus (6)}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{Rumus (7)}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{Rumus (8)}$$

e. Pengujian Model

Model diuji dengan **cross-validation (k=5)** untuk menghindari overfitting dan mengukur generalisasi. Uji coba dilakukan pada **lingkungan simulatif laboratorium komputer**

4. HASIL DAN PEMBAHASAN

Evaluasi Model Naive Bayes

Model klasifikasi dokumen publik berbasis algoritma Naive Bayes dievaluasi menggunakan 100 data uji yang telah dianotasi secara manual ke dalam empat kategori aksesibilitas berdasarkan UU KIP: Berkala, Setiap Saat, Serta Merta, dan Dikecualikan. Evaluasi dilakukan dengan menghitung empat metrik utama, yaitu akurasi, precision, recall, dan F1-score dengan pendekatan rata-rata makro (macro average), yang mengukur performa rata-rata antar semua kategori tanpa memperhitungkan ketidakseimbangan kelas.

Tabel 1. Hasil Evaluasi Model Naive Bayes

Metrik Evaluasi	Skor
Akurasi	0.83
Precision (Macro)	0.83
Recall (Macro)	0.83
F1-Score (Macro)	0.82

Sumber: Hasil pengolahan data dengan 100 dokumen publik menggunakan algoritma Naive Bayes.

Nilai-nilai tersebut menunjukkan bahwa performa model tergolong baik dalam mengenali berbagai kategori dokumen, terutama untuk kelas yang lebih umum seperti Setiap Saat dan Berkala. Meskipun demikian, model mengalami kesulitan dalam mengklasifikasikan dokumen dengan konteks semantik yang lebih kompleks, seperti dokumen Dikecualikan yang mengandung istilah hukum dan konteks sensitif.

Keunggulan utama dari Naive Bayes terletak pada kesederhanaan dan efisiensi komputasi, menjadikannya cocok digunakan dalam sistem dengan sumber daya terbatas. Namun, keterbatasannya terletak pada ketidakmampuannya dalam memahami relasi kontekstual antar kata secara mendalam, sehingga menghasilkan prediksi yang kurang tepat pada dokumen berdensitas tinggi atau dokumen hukum.

Evaluasi Model K-Nearest Neighbors (KNN)

Selain algoritma Naive Bayes, model klasifikasi dokumen juga diuji menggunakan algoritma K-Nearest Neighbors (KNN) dengan nilai k=5. KNN bekerja dengan mencari lima dokumen terdekat, berdasarkan kemiripan vektor fitur dan menentukan kelas mayoritas sebagai hasil prediksi. Algoritma ini dikenal efektif dalam lingkungan dengan struktur data yang jelas dan jumlah fitur yang dapat direpresentasikan secara seimbang.

Evaluasi terhadap model KNN dilakukan pada dataset yang sama dengan 100 dokumen publik, dengan pendekatan yang sama menggunakan metrik akurasi, precision, recall, dan F1 score (macro average). Hasilnya disajikan pada Tabel 2.

Tabel 2. Hasil Evaluasi Model K-Nearest Neighbors (KNN)

Metrik Evaluasi	Skor
Akurasi	0.81
Precision (Macro)	0.80
Recall (Macro)	0.81
F1-Score (Macro)	0.80

Sumber: Hasil pengolahan 100 dokumen publik menggunakan algoritma KNN (k=5).

Hasil evaluasi menunjukkan bahwa performa KNN sedikit lebih rendah dibandingkan dengan Naive Bayes. Hal ini dapat disebabkan oleh sensitivitas KNN terhadap distribusi data dan fitur yang kurang representatif dalam klasifikasi teks, terutama jika tidak didukung oleh teknik seleksi fitur atau reduksi dimensi yang kuat.

Meskipun demikian, KNN memiliki keunggulan dalam menangani kasus-kasus ambiguitas kelas pada dokumen pendek, karena mempertimbangkan konteks lokal dari tetangga terdekat. Namun, performa model ini menurun ketika dokumen memiliki panjang dan variasi kata yang tinggi, atau ketika terdapat noise dalam data.

Evaluasi Model BERT

Model BERT digunakan untuk mengklasifikasikan dokumen publik ke dalam empat kategori aksesibilitas, yaitu: Setiap Saat, Serta Merta, Berkala, dan Dikecualikan. Model ini memanfaatkan kemampuan contextual embedding yang mampu menangkap makna kata dalam konteks kalimat secara dua arah. Dengan demikian, BERT menjadi kandidat kuat untuk menangani dokumen yang mengandung variasi istilah administratif dan hukum.

Model BERT dilatih menggunakan data dokumen publik yang telah direpresentasikan melalui tokenizer pre-trained multilingual BERT. Hasil pelatihan dan evaluasi menunjukkan bahwa model BERT menghasilkan performa yang lebih tinggi dibandingkan dengan model tradisional seperti Naive Bayes dan KNN.

Berikut ini adalah hasil evaluasi model BERT berdasarkan metrik pengukuran standar untuk klasifikasi: Akurasi, Precision, Recall, dan F1-Score.

Tabel 2. Hasil Evaluasi Model K-Nearest Neighbors (KNN)

Metrik Evaluasi	Skor
Akurasi	0.89
Precision (Macro)	0.88
Recall (Macro)	0.89
F1-Score (Macro)	0.88

Sumber: Hasil pengolahan 100 dokumen publik menggunakan algoritma BERT

Kesesuaian Hasil dengan Teori NLP dan Klasifikasi Dokumen

Hasil yang diperoleh dari model BERT menunjukkan konsistensi dengan teori Natural Language Processing (NLP), khususnya dalam klasifikasi teks berbasis representasi kontekstual. Model BERT yang mampu memahami konteks kata dalam kalimat dua arah terbukti unggul dalam mengklasifikasikan dokumen publik yang memiliki variasi istilah administratif dan legal. Hasil evaluasi menunjukkan metrik akurasi yang tinggi, mendekati

89%, dan nilai F1-Score 0.88, yang menegaskan efektivitas pendekatan Transformer dalam menangani klasifikasi berbasis regulasi.

Berdasarkan tujuan penelitian yaitu mengembangkan model klasifikasi dokumen publik berbasis AI/NLP dalam kerangka Manajemen Daur Hidup Data (MDHD). Hasil yang diperoleh menunjukkan bahwa integrasi NLP berbasis BERT ke dalam proses klasifikasi dokumen berhasil menjawab tantangan dalam pengelolaan dokumen informasi publik. Kinerja model dalam simulasi laboratorium yang meniru operasional PPID juga menunjukkan potensi adaptasi tinggi terhadap kebutuhan nyata badan publik.

Implikasi Teoritis dan Praktis

Secara teoritis, penelitian ini memberikan kontribusi pada pengembangan model klasifikasi dokumen berbasis AI/NLP dengan mengadopsi arsitektur BERT dalam konteks sistem informasi publik. Hal ini memperkaya khazanah literatur pada bidang NLP terapan untuk sektor pemerintahan, serta membuktikan efektivitas pendekatan DSR dan Agile dalam proyek pengembangan sistem cerdas berbasis simulasi.

Secara praktis, hasil penelitian ini dapat diimplementasikan oleh instansi pemerintah untuk mempercepat proses klasifikasi dokumen, mengurangi beban kerja manual PPID, dan meningkatkan akurasi layanan informasi publik. Model ini juga dapat diintegrasikan lebih lanjut ke dalam sistem informasi PPID yang sudah berjalan, dengan menambahkan fitur klasifikasi otomatis berbasis AI sebagai decision support system (DSS).

5. KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini berhasil mengembangkan model klasifikasi dokumen publik berbasis AI/NLP dalam kerangka manajemen daur hidup data (MDHD). Model BERT menunjukkan performa terbaik dibandingkan Naive Bayes dan KNN, dengan akurasi sebesar 89%, precision dan recall sebesar 0.88 dan 0.89, serta F1-Score 0.88. Hasil ini menunjukkan efektivitas pendekatan Transformer dalam mengklasifikasikan dokumen publik berdasarkan kategori aksesibilitas sesuai regulasi UU No. 14 Tahun 2008. Model ini berpotensi menjadi solusi cerdas, akuntabel, dan efisien untuk mendukung digitalisasi layanan publik dan agenda nasional seperti RPJMN 2025 dan Asta Cita.

Saran

1. Implementasi sistem sebaiknya dilanjutkan pada lingkungan PPID nyata agar validitas dan keandalan model dapat diuji secara langsung di lapangan.
2. Perluasan dataset dengan melibatkan dokumen dari berbagai badan publik di tingkat pusat dan daerah untuk meningkatkan generalisasi model.
3. Integrasi sistem klasifikasi ini ke dalam platform layanan informasi publik yang telah ada akan memperkuat fungsi decision support system.
4. Pengembangan fitur keamanan dan kebijakan privasi perlu diperhatikan agar proses klasifikasi tidak melanggar prinsip keterbukaan informasi publik.

UCAPAN TERIMA KASIH

Penelitian ini dapat terlaksana atas dukungan sarana prasarana dan sumber dana dari Direktorat Penelitian Pengabdian Kepada Masyarakat dan Publikasi (DPPMP) Uniiversitas Stikubank. Terimakasih atas kerjasama yang baik telah terjalin hingga luaran penelitian berupa artikel ilmiah dapat diwujudkan.

DAFTAR REFERENSI

Aryasatya, R., & Lusiana, V. (2024). Penentuan klustering Indeks Pembangunan Manusia Provinsi Jawa Tengah dengan metode K-Means berbasis web. *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, *8*(1), 155–162. https://doi.org/10.35870/jtik.v8i1.1403

Bennich, A. (2024). The digital imperative: Institutional pressures to digitalise. *Technology in Society*, *76*, 102436. https://doi.org/10.1016/j.techsoc.2023.102436

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186).

Dubey, A., & R, U. S. (2024). Intelligent agriculture system using KNN algorithm. *International Journal of Advanced Research in Science, Communication and Technology*, 52–56. https://doi.org/10.48175/IJARSCT-22512

Felix, C. A., Obiokafor, I. N., & Emenike, A. O. (2023). Sustainable data governance in the era of global data security challenges in Nigeria: A narrative review. *World Journal of Advanced Research and Reviews*, *17*(2), 378–385. https://doi.org/10.30574/wjarr.2023.17.2.0154

Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Republik Indonesia. (2024). *Laporan pelaksanaan evaluasi Sistem Pemerintahan Berbasis Elektronik (SPBE) tahun 2023*.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Linthorst, J., & de Waal, A. (2020). Megatrends and disruptors and their postulated impact on organizations. *Sustainability*, *12*(20), 8740. https://doi.org/10.3390/su12208740

Liu, D. (2023). Improvement of Naive Bayes text classifier based on ensemble technology and feature engineering. In *ICIAAI 2023*, 557–563. https://doi.org/10.2991/978-94-6463-300-9_57

Lusiana, V., Al Amin, I. H., Hartono, B., & Kristianto, T. (2019). Ekstraksi fitur tekstur menggunakan matriks GLCM pada citra dengan variasi arah obyek. *Prosiding SENDI_U*, 978–979.

Muslimin, M., & Lusiana, V. (2023). Analisis sentimen terhadap kenaikan harga bahan pokok menggunakan metode Naive Bayes Classifier. *Jurnal Media Informatika Budidarma*, *7*(3), 1200. https://doi.org/10.30865/mib.v7i3.6418

Pemerintah Republik Indonesia. (2010). *Peraturan Pemerintah No. 61 Tahun 2010 tentang Pelaksanaan Undang-undang No. 14 Tahun 2008 tentang Keterbukaan Informasi Publik*.

Pichiyan, V., Muthulingam, S., Sathar, G., Nalajala, S., Ch, A., & Das, M. N. (2023). Web scraping using natural language processing: Exploiting unstructured text for data extraction and analysis. *Procedia Computer Science*, *230*, 193–202. https://doi.org/10.1016/j.procs.2023.12.074

Prasetyo, A., Darmawan, M., & Moelyana, R. (2019). Analisis dan perancangan tata kelola data sistem pemerintahan berbasis elektronik domain data quality management pada DAMA DMBOK V2 (Studi kasus: Diskominfotik KBB). *E-Proceeding of Engineering*, *6*(2), 7775–7786.

Retnowati, R., Anwar, S. N., & Purwatiningtyas. (2021). Public information management sustainability priority model with a socio-technical approach and Analytic Network Process (ANP) methods (Case study of Salatiga City PPID). *Budapest International Research and Critics Institute (BIRCI) Journal*, *4*(4), 14011–14026. https://www.bircu-journal.com/index.php/birci/article/view/3505

Retnowati, R., Listiyono, H., Anwar, S. N., Studi, P., Informatika, M., Informasi, F. T., & Stikubank, U. (2019). Pengaruh pemanfaatan situs PPIP. *SINTAK*, 289–297.

Retnowati, R., Listiyono, H., Purwatiningtyas, P., Wedaningsih, A. S., & Rahmaziana, L. (2019). Analisis readiness penerapan keterbukaan informasi publik (KIP) dengan pendekatan soft systems methodology (SSM). *Dinamik*, *24*(1), 41–56. https://doi.org/10.35315/dinamik.v24i1.7838

Retnowati, R., Manongga, D. H. F., & Sunarto, H. (2018). Prinsip-prinsip open government data. *CENTIVE*, 25–29.

Retnowati, R., Manongga, D., & Sunarto, H. (2019). Development of sustainability systems for open government data (OGD) management by combining the SHEL model and soft systems methodology analysis. *Journal of Theoretical and Applied Information Technology*, *97*(12).

Retnowati, R., Wahyudi, E. N., & Anis, Y. (2022). Mengukur e-participation masyarakat di era transformasi digital dengan metode Multi Factor Evaluation Process (MFEP). *Jurnal Teknologi dan Sistem Komputer*, *8*(2).

Sofyan, H., Kaswidjanti, W., & Ilmiyah, L. S. (2024). Information Security Index (ISI) 4.2 for information security evaluation (Case study: Sleman Regency Communication and Informatics Office). In *International Conference on Advanced Informatics and Intelligent Information Systems (ICAI3S 2023)*, 188–200. https://doi.org/10.2991/978-94-6463-366-5_18

Sternad Zabukovšek, S., Jordan, S., & Bobek, S. (2023). Managing document management systems' life cycle in relation to an organization's maturity for digital transformation. *Sustainability*, *15*(21), 15212. https://doi.org/10.3390/su152115212

Undang-Undang Republik Indonesia No. 14 Tahun 2008 tentang Keterbukaan Informasi Publik.

Wardhani, W. K., Soewito, B., & Zarlis, M. (2023). Information security evaluation using case study Information Security Index on licensing portal applications. *Journal of Information Systems and Informatics*, *5*(4), 1204–1220. https://doi.org/10.51519/journalisi.v5i4.563