BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

Evaluasi Kinerja Pendekatan Latent Semantic Indexing dan Kombinasi Latent Semantic Indexing - K Nearest Neighbor pada Klasifikasi Dokumen Beban Kerja Dosen

Khairun Nadiah*, Mansur As, Said Iskandar Al Idrus, Mulyono, Zulfahmi Indra

Fakultas Matematika dan Ilmu Pengetahuan Alam, Program Studi Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia
Email: [1,*]khairunnadiah21@gmail.com, [2]asmansur@unimed.ac.id, [3]saidiskandar@unimed.ac.id, [4]mulyono_mat@yahoo.com, [5]zulfahmi.indra@unimed.ac.id
Email Penulis Korespondensi: khairunnadiah21@gmail.com

Abstrak−Dokumen Beban Kerja Dosen (BKD) merupakan dokumen administrasi akademik yang memiliki kategori berdasarkan aktivitas tridharma perguruan tinggi sehingga dapat dimanfaatkan sebagai dataset untuk penelitian klasifikasi dokumen. Karakteristik dokumen BKD yang memiliki variasi istilah akademik serta distribusi kelas yang tidak seimbang menjadi tantangan dalam proses klasifikasi. Penelitian ini bertujuan mengevaluasi dan membandingkan kinerja dua pendekatan klasifikasi berbasis *Latent Semantic Indexing* (LSI), yaitu LSI + *Cosine Similarity* dan LSI + *K-Nearest Neighbor* (KNN). Sebagai pembanding, digunakan TF-IDF + KNN sebagai metode *baseline* untuk menganalisis pengaruh penggunaan LSI terhadap performa klasifikasi. Dataset yang digunakan terdiri atas 352 dokumen BKD Pascasarjana Universitas Negeri Medan yang telah melalui tahapan ekstraksi teks, *preprocessing* (*case folding*, *cleaning, tokenisasi, stopword removal, stemming,* normalisasi, dan penambahan kata kunci BKD), pembobotan TF-IDF, serta reduksi dimensi menggunakan LSI. Evaluasi dilakukan menggunakan *Stratified 3-Fold Cross Validation* dengan metrik *Accuracy, Precision, Recall, F1-Score, ROC-AUC,* dan *Average Precision* (AP). Hasil penelitian menunjukkan bahwa pendekatan LSI + *Cosine Similarity* memberikan performa terbaik dengan *Accuracy* 86,3%, *F1-Weighted* 0,865, *AUC-Macro* 0,965, dan *AP-Macro* 0,887. Sementara itu, LSI + KNN memperoleh *Accuracy* 84,6% dan TF-IDF + KNN sebagai *baseline* memperoleh *Accuracy* 82,6%. Hasil tersebut menunjukkan bahwa representasi semantik menggunakan LSI mampu meningkatkan kualitas klasifikasi dokumen BKD, sedangkan pendekatan berbasis *Cosine Similarity* lebih stabil dibandingkan KNN pada dataset yang memiliki variasi istilah akademik dan distribusi kelas yang tidak seimbang.

Kata Kunci: Beban Kerja Dosen; Klasifikasi Dokumen; Latent Semantic Indexing; Cosine Similarity; K-Nearest Neighbor

Abstract−The Lecturer Workload (BKD) document is an academic administrative document categorized based on higher education's Tridharma (triple dharma) activities, making it highly valuable as a dataset for document classification research. However, the characteristics of BKD documents which feature diverse academic terms and an imbalanced class distribution pose a distinct challenge in the classification process. This study aims to evaluate and compare the performance of two classification approaches based on Latent Semantic Indexing (LSI): LSI + Cosine Similarity and LSI + K-Nearest Neighbor (KNN). As a benchmark, TF-IDF + KNN is utilized as a baseline method to analyze the impact of LSI on classification performance. The dataset consists of 352 Postgraduate BKD documents from Universitas Negeri Medan, which underwent text extraction, preprocessing (case folding, cleaning, tokenization, stopword removal, stemming, normalization, and BKD keyword enrichment), TF-IDF weighting, and dimensionality reduction using LSI. Evaluation was conducted using Stratified 3-Fold Cross Validation with Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Average Precision (AP) as metrics. The results indicate that the LSI + Cosine Similarity approach delivers the best performance, achieving an Accuracy of 86.3%, F1-Weighted of 0.865, AUC-Macro of 0.965, and AP-Macro of 0.887. Meanwhile, LSI + KNN achieved 84.6% Accuracy, and the TF-IDF + KNN baseline reached 82.6% Accuracy. These findings demonstrate that semantic representation using LSI successfully enhances the quality of BKD document classification, while the Cosine Similarity-based approach proves more stable than KNN on datasets characterized by varied academic terminology and imbalanced class distributions.

Keywords: Lecturer Workload; Document Classification; Latent Semantic Indexing; Cosine Similarity; K-Nearest Neighbor

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong digitalisasi berbagai aktivitas di lingkungan perguruan tinggi, termasuk dalam pengelolaan administrasi akademik. Salah satu dokumen administrasi yang memiliki peran penting adalah Beban Kerja Dosen (BKD), yaitu dokumen yang memuat laporan pelaksanaan tridharma perguruan tinggi yang terdiri atas kegiatan pendidikan, penelitian, pengabdian kepada masyarakat, dan unsur penunjang. Dokumen BKD digunakan sebagai dasar evaluasi kinerja dosen sehingga setiap semester perguruan tinggi harus mengelola dokumen dalam jumlah yang terus meningkat seiring bertambahnya aktivitas akademik dosen [1]. Kondisi tersebut mendorong pemanfaatan teknik text mining untuk membantu proses pengelompokan dan pengelolaan dokumen secara lebih efektif.

Dalam penelitian ini, dokumen BKD tidak digunakan untuk mengevaluasi kinerja dosen maupun menilai pelaksanaan tridharma perguruan tinggi. Dokumen BKD dimanfaatkan sebagai dataset klasifikasi karena memiliki kategori yang telah ditentukan berdasarkan aktivitas tridharma. Karakteristik tersebut menjadikan dokumen BKD sesuai digunakan sebagai objek evaluasi metode klasifikasi, karena setiap dokumen telah memiliki label yang dapat dijadikan *ground truth.* Selain itu, dokumen BKD memiliki variasi istilah akademik yang cukup tinggi sehingga menjadi media yang baik untuk menguji kemampuan suatu metode dalam merepresentasikan informasi tekstual dan mengelompokkan dokumen ke dalam kategori yang benar.

Berbagai metode telah dikembangkan untuk meningkatkan performa klasifikasi dokumen teks. Salah satu metode representasi dokumen yang banyak digunakan adalah *Latent Semantic Indexing* (LSI). LSI memanfaatkan teknik *Singular Value Decomposition* (SVD) untuk mereduksi dimensi data sehingga hubungan semantik antar istilah dapat direpresentasikan dengan lebih baik dibandingkan pendekatan yang hanya mengandalkan frekuensi kemunculan kata [2].

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

Representasi tersebut memungkinkan dokumen yang memiliki konteks serupa tetapi menggunakan istilah yang berbeda berada pada ruang semantik yang lebih berdekatan.

Representasi dokumen menggunakan LSI umumnya dipadukan dengan dua pendekatan yang berbeda. Pendekatan pertama menggunakan *Cosine Similarity* untuk menghitung tingkat kemiripan antara dokumen uji dengan representasi kategori dokumen. Pendekatan ini relatif sederhana karena klasifikasi dilakukan berdasarkan tingkat kedekatan semantik dokumen terhadap representasi kelas. Pendekatan kedua mengombinasikan LSI dengan algoritma *K-Nearest Neighbor* (KNN), yaitu metode klasifikasi yang menentukan kategori dokumen berdasarkan mayoritas tetangga terdekat pada ruang fitur hasil reduksi dimensi. Kedua pendekatan tersebut sama-sama memanfaatkan LSI sebagai representasi dokumen, tetapi menggunakan mekanisme klasifikasi yang berbeda sehingga berpotensi menghasilkan performa yang berbeda pula.

Sejumlah penelitian menunjukkan bahwa LSI, *Cosine Similarity*, maupun KNN mampu memberikan hasil yang baik pada berbagai permasalahan klasifikasi dokumen, namun masing-masing memiliki fokus dan konteks yang berbeda dari penelitian ini. Istighfarizky et al. (2022) [3] menerapkan kombinasi LSI dan KNN pada klasifikasi abstrak ilmiah berbahasa Indonesia dan menunjukkan bahwa penggunaan LSI mampu meningkatkan kualitas representasi dokumen sebelum proses klasifikasi dilakukan dengan akurasi sebesar 84,3%; penelitian tersebut hanya menguji satu jalur pemodelan (LSI-KNN) tanpa membandingkannya dengan pendekatan LSI-*Cosine Similarity* maupun *baseline* TF-IDF, dan objek penelitiannya berupa abstrak ilmiah, bukan dokumen administratif perguruan tinggi. Murniati (2021) [4] memanfaatkan LSI bersama *Cosine Similarity* pada analisis sentimen aplikasi *mobile banking* dan memperoleh akurasi sebesar 81,67%, membuktikan bahwa representasi semantik mampu meningkatkan pengukuran kemiripan antar dokumen, penelitian ini diterapkan pada ulasan pengguna aplikasi, bukan dokumen administrasi akademik, dan tidak melibatkan perbandingan dengan pendekatan KNN maupun *baseline* TF-IDF.

Penelitian lain yang dilakukan oleh Rivaldi et al. (2024) [5] menunjukkan bahwa kombinasi LSI dan KNN menghasilkan performa yang lebih stabil pada dataset berukuran besar karena proses reduksi dimensi mampu mengurangi *noise* yang memengaruhi proses klasifikasi, penelitian tersebut berfokus pada skalabilitas dataset umum berukuran besar. Ajijah dan Kurniawan (2023) [6] menunjukkan bahwa pendekatan LSI-*Cosine Similarity* efektif digunakan untuk mendeteksi kemiripan semantik pada dokumen administrasi pendidikan tinggi, meskipun objeknya cukup dekat dengan penelitian ini, penelitian tersebut tidak membandingkan LSI-*Cosine Similarity* secara langsung dengan LSI-KNN maupun TF-IDF+KNN pada dataset dan skenario evaluasi yang identik. Sementara itu, Kosasih dan Alberto (2021) [7] membuktikan bahwa kombinasi TF-IDF dan KNN mampu menghasilkan performa klasifikasi yang baik pada analisis sentimen dengan akurasi sebesar 79,33%, namun tanpa menerapkan reduksi dimensi sama sekali sehingga kontribusi LSI terhadap peningkatan performa tidak dapat diukur dari penelitian tersebut. Mehta et al. (2023) [8] menunjukkan bahwa algoritma KNN tetap memiliki kemampuan yang baik dalam klasifikasi dokumen berbasis teks pada domain keamanan perangkat lunak, tetapi penelitian tersebut berada pada domain yang sangat berbeda (keamanan siber) dan tidak melibatkan representasi semantik berbasis LSI maupun perbandingan dengan *Cosine Similarity*.

Berdasarkan kondisi tersebut, *research gap* dalam penelitian ini bukan terletak pada belum adanya metode klasifikasi dokumen, melainkan pada belum adanya evaluasi yang membandingkan secara langsung kinerja pendekatan LSI-*Cosine Similarity* dan kombinasi LSI-KNN pada klasifikasi dokumen BKD. Padahal, kedua pendekatan tersebut sama-sama memanfaatkan representasi semantik menggunakan LSI, tetapi memiliki mekanisme penentuan kelas yang berbeda. LSI-*Cosine Similarity* menentukan kategori berdasarkan tingkat kemiripan semantik terhadap representasi kelas, sedangkan LSI-KNN menentukan kategori berdasarkan mayoritas tetangga terdekat pada ruang semantik hasil reduksi dimensi. Perbedaan mekanisme tersebut memungkinkan masing-masing pendekatan memiliki karakteristik performa yang berbeda ketika diterapkan pada dataset yang sama.

Penelitian ini juga menggunakan TF-IDF + KNN sebagai metode *baseline* untuk memberikan gambaran mengenai pengaruh penggunaan LSI sebagai teknik reduksi dimensi terhadap performa klasifikasi. Dengan adanya metode *baseline*, peningkatan maupun penurunan performa yang diperoleh dari pendekatan berbasis LSI dapat dianalisis secara lebih objektif. Seluruh pendekatan dievaluasi menggunakan dataset dokumen BKD Pascasarjana Universitas Negeri Medan, tahapan *preprocessing* yang sama, serta metrik evaluasi yang identik sehingga hasil perbandingan benar-benar mencerminkan perbedaan kinerja metode, bukan dipengaruhi oleh perbedaan data maupun skenario pengujian.

Kebaruan (*novelty*) penelitian ini terletak pada evaluasi komparatif dua pendekatan klasifikasi berbasis LSI, yaitu LSI-*Cosine Similarity* dan LSI-KNN, menggunakan dataset dokumen BKD yang memiliki karakteristik istilah akademik berbahasa Indonesia. Berbeda dengan penelitian sebelumnya yang umumnya menggunakan dokumen berita, abstrak ilmiah, ulasan produk, maupun analisis sentimen, penelitian ini memanfaatkan dokumen administratif perguruan tinggi sebagai objek penelitian. Selain itu, seluruh metode dibandingkan menggunakan dataset dan skenario evaluasi yang sama sehingga hasil penelitian mampu memberikan gambaran yang lebih objektif mengenai efektivitas masing-masing pendekatan.

Penelitian ini memberikan beberapa kontribusi. Pertama, penelitian ini menyajikan perbandingan secara langsung antara metode LSI-*Cosine Similarity* dan LSI-KNN dalam klasifikasi dokumen BKD, yang hingga saat ini masih jarang ditemukan pada penelitian sebelumnya. Kedua, penelitian ini menggunakan TF-IDF + KNN sebagai metode *baseline* sehingga pengaruh penerapan reduksi dimensi menggunakan LSI terhadap kinerja klasifikasi dapat dievaluasi secara lebih objektif. Ketiga, penelitian ini memperluas penerapan metode klasifikasi berbasis LSI pada dokumen administratif akademik berbahasa Indonesia, yang masih relatif sedikit diteliti dibandingkan dengan penerapannya pada dokumen berita, ulasan produk, maupun analisis sentimen.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

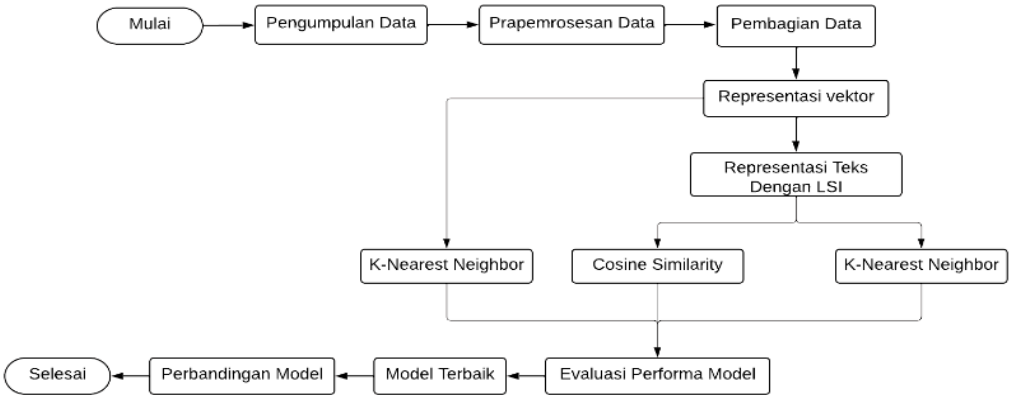
Berdasarkan uraian tersebut, penelitian ini bertujuan mengevaluasi dan membandingkan kinerja pendekatan LSI-*Cosine Similarity* dan kombinasi LSI-KNN pada klasifikasi dokumen Beban Kerja Dosen. Sebagai pembanding, penelitian ini juga menggunakan TF-IDF + KNN sebagai metode *baseline* untuk menganalisis pengaruh penggunaan LSI terhadap performa klasifikasi. Evaluasi dilakukan menggunakan *metrik Accuracy, Precision, Recall,* dan *F1-Score* sehingga diperoleh informasi mengenai pendekatan yang memberikan performa terbaik pada klasifikasi dokumen BKD. Hasil penelitian diharapkan dapat menjadi referensi bagi penelitian selanjutnya dalam memilih pendekatan klasifikasi dokumen berbasis *text mining,* khususnya pada dokumen administratif di lingkungan perguruan tinggi.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan metode eksperimen komparatif untuk mengevaluasi dan membandingkan kinerja dua pendekatan klasifikasi dokumen, yaitu *Latent Semantic Indexing* (LSI) dengan *Cosine Similarity* dan kombinasi LSI dengan *K-Nearest Neighbor* (KNN). Sebagai pembanding (*baseline*), penelitian ini juga menerapkan metode TF-IDF + KNN tanpa reduksi dimensi. Seluruh pendekatan diuji menggunakan dataset, tahapan *preprocessing*, serta skenario evaluasi yang sama sehingga perbedaan performa yang diperoleh mencerminkan pengaruh metode yang digunakan.

Tahapan penelitian diawali dengan pengumpulan dokumen BKD, dilanjutkan dengan *preprocessing*, pembagian data, pembentukan representasi dokumen menggunakan TF-IDF dan LSI, proses klasifikasi, serta evaluasi model menggunakan metrik *Accuracy, Precision, Recall, F1-Score, ROC-AUC, Average Precision* (AP), *Learning Curve,* dan *Reliability Curve*. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini dikumpulkan dari dokumen Beban Kerja Dosen (BKD) yang diperoleh dari lingkungan Pascasarjana Universitas Negeri Medan yang mencakup dokumen dari berbagai kategori aktivitas tridharma perguruan tinggi. Dokumen BKD dipilih karena telah memiliki kategori berdasarkan aktivitas tridharma perguruan tinggi sehingga dapat digunakan sebagai *ground truth* dalam proses pelatihan dan evaluasi model klasifikasi. Penelitian ini tidak bertujuan untuk mengevaluasi isi maupun kualitas dokumen BKD, melainkan memanfaatkan dokumen tersebut sebagai data untuk membandingkan performa pendekatan klasifikasi berbasis *text mining.*

2.3 Pra-pemrosesan Data

Tahap pra-pemrosesan (*preprocessing*) bertujuan menyiapkan dokumen agar dapat diproses oleh algoritma klasifikasi. Proses ini dilakukan untuk menghilangkan informasi yang tidak relevan, menyeragamkan bentuk kata, serta menghasilkan representasi teks yang lebih konsisten sehingga mampu meningkatkan kualitas fitur yang digunakan pada proses klasifikasi. Tahapan *preprocessing* yang dilakukan meliputi:

a. Ekstraksi Teks: Dokumen BKD yang masih berbentuk PDF dikonversi menjadi teks menggunakan mekanisme *fallback* bertingkat. Proses ekstraksi diawali menggunakan *PyPDF2,* kemudian *pdfminer* apabila teks tidak dapat diekstraksi secara optimal. Selanjutnya, apabila dokumen berupa hasil pemindaian (*scanned document*), proses *Optical Character Recognition* (OCR) dilakukan menggunakan *Tesseract* OCR sehingga seluruh dokumen dapat diproses dalam bentuk teks.
b. *Case Folding*: Tahap *Case Folding* adalah proses pengubahan seluruh karakter teks ke dalam bentuk huruf kecil, yang juga mencakup penghapusan tanda baca serta angka [9].
c. *Cleaning*: Tahap *Cleaning* data adalah proses persiapan data untuk analisis dengan cara menghilangkan atau memodifikasi data yang tidak akurat, tidak lengkap, tidak relevan. Pembersihan data bertujuan untuk menghilangkan *noise* serta elemen data yang tidak relevan [10].
d. *Tokenisasi*: *Tokenisasi* merupakan proses pemecahan teks menjadi serangkaian token yang kemudian dikodekan ke dalam format yang dapat ditafsirkan mesin [11].

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

e. *Stopword Removal*: Kata-kata umum yang memiliki frekuensi kemunculan tinggi tetapi tidak memberikan informasi penting, seperti dan, yang, di, ke, serta kata hubung lainnya, dihapus menggunakan kamus *stopword* bahasa Indonesia. Tujuannya adalah untuk membersihkan token yang dihasilkan dari proses sebelumnya dengan menghapus kata-kata umum yang dianggap tidak penting [12].
f. *Stemming*: Stemming merupakan salah satu tahapan prapemrosesan teks yang bertujuan mengubah kata-kata berimbuhan menjadi bentuk dasarnya [13]. Tahap *stemming* merupakan proses yang menghasilkan bentuk dasar dari akar kata/kata dasar. Dengan kata lain, proses ini mereduksi kata-kata menjadi kata dasarnya.
g. Normalisasi: Normalisasi teks diartikan sebagai proses mengubah teks tidak standar menjadi bentuk standar melalui perbaikan kesalahan ejaan, tata bahasa, dan kesalahan lainnya. Akurasi dalam pengumpulan kata-kata yang benar dan konsisten untuk analisis lebih lanjut dapat ditingkatkan melalui hasil dari normalisasi ini [14].
h. Penambahan Kata Kunci BKD: Tahap terakhir merupakan penambahan kata kunci yang berkaitan dengan kategori BKD. Kata kunci tersebut disisipkan ke dalam dokumen apabila ditemukan istilah yang relevan, seperti RPS, HKI, pengelola jurnal, pembicara, dan istilah akademik lainnya. Penambahan kata kunci bertujuan memperkuat representasi fitur sehingga hubungan antara isi dokumen dan kategori BKD dapat dikenali dengan lebih baik oleh model klasifikasi.

2.4 Representasi Dokumen

Dokumen direpresentasikan ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Representasi ini bertujuan mengubah dokumen teks menjadi vektor sehingga dapat diproses oleh algoritma klasifikasi. TF-IDF memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya pada suatu dokumen (*term frequency*) dan tingkat kepentingannya terhadap keseluruhan koleksi dokumen (*inverse document frequency*). merepresentasikan kata tunggal (*unigram*) maupun pasangan kata (*bigram*) yang sering muncul pada dokumen BKD. Representasi TF-IDF selanjutnya digunakan pada tiga jalur pemodelan yang berbeda. Pada jalur pertama dan kedua, representasi TF-IDF direduksi dimensinya menggunakan metode *Latent Semantic Indexing* (LSI), sedangkan pada jalur ketiga representasi TF-IDF digunakan secara langsung sebagai *baseline* tanpa reduksi dimensi.

LSI merupakan metode reduksi dimensi yang memanfaatkan teknik *Singular Value Decomposition* (SVD) untuk memperoleh representasi dokumen pada ruang semantik berdimensi lebih rendah. Melalui proses ini, hubungan *laten* antar istilah dapat dipertahankan sekaligus mengurangi *noise* yang berasal dari tingginya jumlah fitur pada representasi TF-IDF. Secara matematis, proses dekomposisi matriks pada LSI dinyatakan menggunakan Persamaan (1).

$$A = U \sum V^T \tag{1}$$

Keterangan: Pada Persamaan (1), A merupakan matriks masukan (*input matrix*) berukuran $m \times n$, U merupakan matriks hubungan antara kata (*words*) dan konsep (*concept*) berukuran $m \times n$, $\sum$ merupakan matriks diagonal yang berisi nilai singular (*singular values)* berukuran $n \times n$, sedangkan V^T merupakan matriks transpos yang menunjukkan hubungan antara kalimat (*sentences*) dan konsep (*concept*) berukuran $n \times n$.

2.5 Metode Klasifikasi

Penelitian ini membandingkan dua pendekatan klasifikasi berbasis LSI, yaitu LSI-*Cosine Similarity* dan kombinasi LSI-KNN. Sebagai pembanding (*baseline*), digunakan metode TF-IDF + KNN tanpa reduksi dimensi. Ketiga pendekatan menggunakan *dataset* dan tahapan *preprocessing* yang sama sehingga perbedaan performa hanya dipengaruhi oleh metode klasifikasi yang digunakan.

2.5.1 LSI + *Cosine Similarity*

Pada pendekatan pertama, dokumen direpresentasikan menggunakan TF-IDF kemudian direduksi dimensinya dengan LSI sehingga diperoleh representasi semantik yang lebih ringkas. Proses klasifikasi dilakukan dengan menghitung tingkat kemiripan (*similarity*) antara dokumen uji dan *centroid* masing-masing kategori menggunakan *Cosine Similarity*. Metode ini diartikan sebagai pendekatan untuk menghitung tingkat kemiripan antar dua objek. Nilai *cosinus* sudut antara dua vektor dihitung oleh metode ini guna mengukur kemiripan antar dua dokumen dengan menggambarkan kesamaan antara vektor *query* dan vektor dokumen yang menghasilkan sudut *cosinus* di antara dua vektor tersebut [15]. Dokumen akan diklasifikasikan ke dalam kategori yang memiliki nilai kemiripan tertinggi. Nilai *Cosine Similarity* dihitung menggunakan Persamaan (2).

$$\cos(\theta) = \frac{A.B}{\|A\|.\|B\|} \tag{2}$$

Keterangan: Pada Persamaan (2), A merupakan vektor pertama, B merupakan vektor kedua yang akan dibandingkan tingkat kemiripannya, $A.B$ merupakan hasil perkalian titik (*dot product*) antara vektor A dan vektor B, $\|A\|$ merupakan panjang vektor A, $\|B\|$ merupakan panjang vektor B, dan $\|A\|.\|B\|$merupakan hasil perkalian antara panjang vektor A dan panjang vektor B.

2.5.2 LSI + *K-Nearest Neighbor* (KNN)

Pendekatan kedua menggunakan hasil transformasi LSI sebagai masukan bagi algoritma KNN. KNN menentukan kategori dokumen berdasarkan mayoritas kelas dari sejumlah tetangga terdekat pada ruang semantik hasil reduksi dimensi. Jarak antar dokumen dihitung menggunakan metrik jarak, kemudian sejumlah tetangga terdekat digunakan

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

sebagai dasar pengambilan keputusan klasifikasi. Nilai parameter (k), fungsi bobot (*weights*), serta metrik jarak dioptimalkan menggunakan *GridSearchCV* sehingga diperoleh konfigurasi terbaik berdasarkan nilai *F1-Weighted*. Algoritma ini termasuk dalam kategori *supervised learning* dan digunakan untuk tugas klasifikasi serta regresi dengan cara memprediksi label suatu data berdasarkan mayoritas label tetangga terdekat [16]. Objek diklasifikasikan oleh KNN berdasarkan kedekatannya dengan data pelatihan yang sudah ada [17]. Secara umum, jarak *Euclidean* dinyatakan menggunakan Persamaan (3).

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

Keterangan: Pada Persamaan (3), $\mathrm{d}(\mathrm{x},\mathrm{y})$ merupakan jarak antara dua titik x dan y; x_i merupakan nilai fitur ke-i dari titik x, dan y_i merupakan nilai fitur ke-i pada titik y.

2.5.1 TF- IDF + KNN (*Baseline*)

Sebagai pembanding, penelitian ini juga menerapkan algoritma KNN secara langsung pada representasi TF-IDF tanpa melalui proses reduksi dimensi LSI. Pendekatan ini digunakan sebagai metode *baseline* untuk mengevaluasi pengaruh penggunaan LSI terhadap performa klasifikasi. Seluruh parameter KNN pada metode *baseline* dioptimalkan menggunakan *GridSearchCV* dengan ruang pencarian parameter yang sama seperti pada model LSI-KNN. Dengan demikian, perbedaan hasil yang diperoleh terutama disebabkan oleh penggunaan representasi fitur, yaitu ruang vektor TF-IDF berdimensi tinggi dibandingkan ruang semantik hasil reduksi LSI.

2.6 Pengujian Model

Tahap pengujian dilakukan untuk mengevaluasi dan membandingkan kinerja tiga pendekatan klasifikasi, yaitu LSI-*Cosine Similarity*, LSI-KNN, dan TF-IDF + KNN sebagai *baseline.* Proses validasi model menggunakan metode *Stratified K-Fold Cross Validation.* Metode ini dipilih karena mampu mempertahankan proporsi setiap kategori pada setiap lipatan (*fold*), sehingga distribusi data latih dan data uji tetap seimbang. Jumlah *fold* disesuaikan dengan jumlah data pada kelas minoritas untuk menghindari terjadinya pembagian data yang tidak representatif.

Selain itu, penelitian ini menerapkan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) pada data pelatihan untuk mengatasi ketidakseimbangan jumlah dokumen antar kategori. SMOTE bekerja dengan membangkitkan sampel sintetis pada kelas minoritas sehingga distribusi data menjadi lebih proporsional. Penerapan SMOTE hanya dilakukan pada data latih agar tidak menyebabkan data *leakage* pada proses evaluasi model. Setelah proses pelatihan selesai, ketiga model dievaluasi menggunakan beberapa metrik performa klasifikasi.

2.6.1 *Accuracy*

Accuracy menunjukkan proporsi keseluruhan dokumen yang berhasil diklasifikasikan dengan benar terhadap seluruh data pengujian. Untuk mengukur keakuratan masing-masing metode maka digunakan penghitungan akurasi, di mana semakin besar nilai akurasi maka metode semakin baik [18].

2.6.2 *Precision*

Precision menunjukkan tingkat ketepatan model dalam memberikan prediksi terhadap suatu kategori. Presisi merupakan rasio antara jumlah data benar bernilai positif dengan hasil jumlah data benar bernilai positif dan data salah bernilai positif [18]. Semakin tinggi nilai precision, semakin sedikit dokumen yang salah diklasifikasikan ke dalam suatu kategori.

2.6.3 *Recall*

Recall menunjukkan kemampuan model dalam menemukan seluruh dokumen yang benar-benar termasuk pada suatu kategori. *Recall* merupakan rasio antara jumlah data benar bernilai positif dengan hasil jumlah data benar yang bernilai positif dan data salah bernilai negatif [18]. Nilai *Recall* yang tinggi menunjukkan bahwa model mampu mengenali sebagian besar dokumen yang memang berasal dari kategori tersebut.

2.6.4 *F1-Score*

F1-Score merupakan rata-rata harmonis antara *Precision* dan *Recall*. Metrik ini digunakan karena memberikan keseimbangan antara ketepatan dan kelengkapan prediksi, terutama pada dataset yang memiliki distribusi kelas tidak seimbang. *F1-score* adalah metrik pengukuran yang menggabungkan presisi dan recall untuk menghasilkan pengukuran tunggal yang merepresentasikan keefektifan suatu klasifikasi. *F1-score* didefinisikan sebagai rata-rata dari presisi dan *recall* [19]. Pada penelitian ini, *F1-Weighted* digunakan sebagai indikator utama dalam membandingkan performa model karena memperhitungkan proporsi jumlah dokumen pada setiap kategori.

2.6.5 ROC-AUC

Receiver Operating Characteristic - Area Under Curve (ROC-AUC) digunakan untuk mengevaluasi kemampuan model dalam membedakan setiap kategori dokumen. Kurva ROC menggambarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai nilai ambang (*threshold*). Semakin besar nilai AUC, semakin baik kemampuan model dalam membedakan antar kategori dokumen. *Analisis Receiver Operating Characteristic* (ROC) serta

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

perhitungan *Area Under the Curve* (AUC) diterapkan untuk menilai kemampuan model dalam membedakan kelas positif dan negatif secara keseluruhan [20].

2.6.6 *Average Precision* (AP)

Kurva *Precision-Recall* digunakan sebagai pelengkap ROC-AUC, khususnya pada dataset yang memiliki distribusi kelas tidak seimbang. Luas area di bawah kurva tersebut dinyatakan sebagai *Average Precision* (AP). Pada penelitian ini digunakan nilai *AP-Macro* untuk menggambarkan performa rata-rata model pada seluruh kategori dokumen BKD.

2.6.7 *Learning Curve*

Learning Curve digunakan untuk mengevaluasi kemampuan generalisasi model berdasarkan perubahan performa terhadap jumlah data pelatihan. Kurva ini menunjukkan hubungan antara ukuran data latih dengan nilai performa pada data pelatihan dan data validasi. Apakah model mengalami *overfitting* atau *underfitting,* dan bagaimana kinerja model meningkat dengan penambahan data pelatihan, dapat dipantau dengan memplot *learning curve* [21].

2.6.8 Analisis *Overfitting*

Analisis *overfitting* dilakukan dengan membandingkan performa model pada data pelatihan dan data validasi selama proses pelatihan. *Overfitting* didefinisikan sebagai kondisi di mana model menunjukkan performa yang sangat bagus pada tahap pelatihan, namun menghasilkan kinerja yang buruk pada tahap pengujian. Kondisi ini terjadi karena model cenderung melakukan hafalan terhadap data latih, termasuk *noise* atau informasi yang tidak relevan yang terdapat di dalamnya [22].

2.6.9 *Reliability Curve*

Reliability Curve digunakan untuk mengevaluasi kualitas probabilitas prediksi yang dihasilkan model. Kurva ini membandingkan probabilitas prediksi dengan probabilitas aktual sehingga dapat diketahui apakah model menghasilkan probabilitas yang terkalibrasi dengan baik. Model dengan kurva yang mendekati garis diagonal menunjukkan bahwa probabilitas prediksi yang dihasilkan memiliki tingkat kepercayaan yang baik terhadap kondisi sebenarnya.

Berdasarkan seluruh metrik evaluasi tersebut, dilakukan analisis komparatif terhadap ketiga pendekatan klasifikasi untuk menentukan metode yang memberikan performa terbaik pada klasifikasi dokumen Beban Kerja Dosen. Hasil evaluasi selanjutnya menjadi dasar dalam pembahasan mengenai efektivitas penggunaan LSI sebagai teknik reduksi dimensi serta perbandingan kinerja antara pendekatan LSI-*Cosine Similarity*, LSI-KNN, dan TF-IDF + KNN sebagai *baseline.* Hasil prediksi atau simulasi yang dihasilkan dapat dipercaya dan dijadikan dasar dalam proses pengambilan keputusan [23].

3. HASIL DAN PEMBAHASAN

3.1 Hasil

3.1.1 Deskripsi dataset

Penelitian ini menggunakan dataset berupa 352 dokumen Beban Kerja Dosen (BKD) yang diperoleh dari lingkungan Pascasarjana Universitas Negeri Medan. Distribusi data menunjukkan adanya ketidakseimbangan jumlah dokumen antar kategori sehingga berpotensi memengaruhi performa model klasifikasi (Tabel 1).

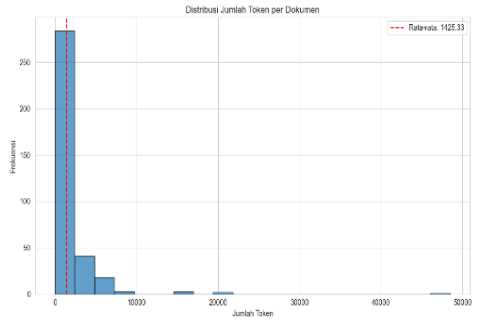
Tabel 1. Distribusi dataset Dokumen BKD

Nama	Nomor	Field
1	Pendidikan – Bahan Ajar	13
2	Pendidikan - Pengajaran	15
3	Pendidikan – Pengujian Mahasiswa	71
4	Penelitian – Tugas Tambahan	7
5	Penelitian – Paten HKI	52
6	Penelitian – Penelitian	11
7	Penelitian – Publikasi Karya	59
8	Pengabdian – Pembicara	53
9	Pengabdian – Pengabdian	35
10	Penunjang – Pengelola Jurnal	7
11	Penunjang –Organisasi Profesi	14
12	Penunjang – Penghargaan	3
13	Penunjang – Penunjang Lain	12
Total	352	

Selain ketidakseimbangan kelas, karakteristik dokumen juga menunjukkan variasi yang cukup tinggi. Dokumen BKD berasal dari berbagai jenis aktivitas akademik, seperti laporan pengajaran, artikel ilmiah, sertifikat paten, surat tugas,

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

piagam penghargaan, dokumen seminar, hingga laporan pengabdian kepada masyarakat. Variasi jenis dokumen tersebut menyebabkan perbedaan struktur, panjang teks, serta penggunaan istilah yang cukup beragam. Hal ini dapat di lihat pada Gambar 2.



Gambar 2. Distribusi Jumlah Token Per Dokumen

Untuk mengatasi ketidakseimbangan data, teknik SMOTE diterapkan pada model LSI + KNN, sedangkan model LSI + *Cosine Similarity* tidak menggunakan SMOTE karena berpotensi menggeser posisi *centroid* kelas.

3.1.2 Pra-pemrosesan Data

Tahap pra-pemrosesan bertujuan untuk mengubah dokumen mentah menjadi data teks yang bersih, konsisten, dan siap direpresentasikan ke dalam bentuk numerik. Tahapan ini sangat penting karena kualitas data masukan sangat memengaruhi performa model klasifikasi.

Tahap pertama adalah ekstraksi teks dari dokumen PDF. Dokumen digital diproses menggunakan pustaka *PyPDF2*, sedangkan dokumen hasil pemindaian (*scan*) diekstraksi menggunakan teknologi *Optical Character Recognition* (OCR) *Tesseract*. Kombinasi kedua metode tersebut memungkinkan seluruh dokumen dikonversi menjadi teks sehingga dapat diproses lebih lanjut. Selanjutnya dilakukan *case folding*, yaitu mengubah seluruh karakter menjadi huruf kecil (*lowercase*). Tahap berikutnya adalah *cleaning*, yang meliputi penghapusan tanda baca, karakter khusus, simbol, dan elemen-elemen lain yang tidak memiliki kontribusi terhadap proses klasifikasi. Selain itu dilakukan pula normalisasi penulisan agar istilah yang memiliki bentuk berbeda dapat direpresentasikan secara seragam. Setelah proses pembersihan selesai, teks dipisahkan menjadi kumpulan token menggunakan proses tokenisasi. Tahap selanjutnya adalah penghapusan *stopword* menggunakan daftar *stopword* bahasa Indonesia. Kata-kata umum seperti dan, yang, di, ke, serta kata-kata lain yang tidak memiliki makna diskriminatif dihapus sehingga hanya tersisa kata-kata penting yang merepresentasikan isi dokumen. Selanjutnya dilakukan proses *stemming* untuk mengubah kata berimbuhan menjadi bentuk dasarnya sehingga variasi kata yang memiliki akar sama dapat direpresentasikan sebagai satu fitur. Penelitian ini menerapkan penambahan kata kunci berbasis aturan (*rule-based keyword enrichment*). Sistem mengenali kata-kata tertentu yang berkaitan dengan kategori BKD. Ketika kata-kata tersebut ditemukan, sistem menambahkan token kategori yang sesuai sehingga informasi semantik dokumen menjadi lebih jelas. Pendekatan ini membantu meningkatkan representasi konsep *laten* pada dokumen BKD yang memiliki istilah akademik yang beragam.

3.1.3 Representasi Dokumen Menggunakan TF-IDF

Setelah proses pra-pemrosesan selesai dilakukan, seluruh dokumen direpresentasikan ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Metode ini digunakan untuk mengukur tingkat kepentingan suatu kata terhadap sebuah dokumen dengan mempertimbangkan frekuensi kemunculan kata pada dokumen tersebut maupun pada seluruh koleksi dokumen. Pembobotan TF-IDF memberikan bobot tinggi kepada kata yang sering muncul pada suatu dokumen namun jarang ditemukan pada dokumen lain. Sebaliknya, kata-kata yang muncul hampir di seluruh dokumen akan memperoleh bobot yang rendah karena dianggap kurang mampu membedakan suatu kategori dokumen. Pada penelitian ini, representasi TF-IDF dibangun menggunakan unigram dan bigram dengan batas maksimum fitur sebanyak 50.000 serta nilai minimum *document frequency* (min_df) sebesar 2. Pengaturan tersebut bertujuan untuk menghilangkan kata-kata yang sangat jarang muncul sekaligus mempertahankan informasi penting yang relevan terhadap proses klasifikasi. Hasil transformasi menghasilkan matriks TF-IDF berdimensi tinggi yang bersifat sparse, yaitu sebagian besar elemennya bernilai nol. Kondisi tersebut umum terjadi pada representasi teks karena setiap dokumen hanya mengandung sebagian kecil dari keseluruhan kosakata yang terdapat dalam korpus. Meskipun TF-IDF mampu merepresentasikan tingkat kepentingan kata dengan baik, tingginya dimensi fitur dapat meningkatkan kompleksitas komputasi sekaligus menimbulkan *fenomena curse of dimensionality*, khususnya pada algoritma berbasis jarak seperti *K-Nearest Neighbor.*

3.1.4 Reduksi Dimensi Menggunakan *Latent Semantic Indexing*

Untuk mengatasi tingginya dimensi ruang fitur TF-IDF, penelitian ini menerapkan *Latent Semantic Indexing* (LSI) yang diimplementasikan menggunakan *Truncated Singular Value Decomposition* (SVD). LSI bertujuan memproyeksikan dokumen ke dalam ruang fitur berdimensi lebih rendah sehingga hubungan semantik antar kata dapat direpresentasikan

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

secara lebih efektif. Melalui proses dekomposisi matriks, TF-IDF diubah menjadi tiga matriks utama, yaitu matriks U, Σ, dan V^T. Matriks tersebut merepresentasikan hubungan antara dokumen, konsep laten, serta istilah yang terdapat pada korpus. Pada penelitian ini jumlah komponen laten ditetapkan sebanyak 300 komponen, sehingga setiap dokumen direpresentasikan dalam ruang semantik yang jauh lebih ringkas dibandingkan ruang fitur TF-IDF awal. Representasi hasil LSI selanjutnya digunakan sebagai masukan pada dua pendekatan klasifikasi yang dibandingkan dalam penelitian ini, yaitu LSI + *Cosine Similarity* dan LSI + *K-Nearest Neighbor* (KNN).

3.1.5 Pembagian Data

Evaluasi model dilakukan menggunakan kombinasi metode *hold-out* dan *3-Fold Cross Validation*. Seluruh dataset dibagi dengan rasio 80:20, di mana sebanyak 80% data digunakan sebagai data pelatihan dan 20% sisanya digunakan sebagai data pengujian akhir. Pada data pelatihan diterapkan *Stratified 3-Fold Cross Validation* sehingga proporsi setiap kelas tetap terjaga pada setiap *fold*. Pendekatan ini dipilih untuk memperoleh estimasi performa model yang lebih stabil sekaligus meminimalkan risiko *overfitting* akibat ukuran dataset yang relatif terbatas. Khusus pada model LSI + KNN dilakukan proses optimasi parameter menggunakan *GridSearchCV* dengan beberapa variasi nilai K, bobot tetangga, dan jenis metrik jarak. Berdasarkan hasil validasi silang, parameter terbaik diperoleh pada K = 5 sehingga digunakan pada proses evaluasi akhir.

3.1.6 Pengujian Model

Penelitian ini membandingkan tiga pendekatan klasifikasi, yaitu LSI + *Cosine Similarity*, LSI + KNN (K = 5), dan TF-IDF + KNN sebagai model pembanding (*baseline*). Evaluasi dilakukan menggunakan metrik *Accuracy, F1-Weighted, AUC-Macro*, dan *Average Precision* (*AP-Macro*). Hasil ringkasan evaluasi ketiga model disajikan pada Tabel 2.

Tabel 2. Ringkasan Hasil Evaluasi Model

Metrik	LSI + *Cosine Similarity*	LSI + KNN (K=5)	TF-IDF + KNN (K=5)
Accuracy	0,863	0,846	0,826
F1-Weighted	0,865	0,848	0,848
AUC-Macro	0,965	0,930	0,927
AP-Macro	0,887	0,766	0,783

Hasil pengujian menunjukkan bahwa model LSI + *Cosine Similarity* memberikan performa terbaik dengan nilai *Accuracy* sebesar 86,3%, *F1-Weighted* sebesar 0,865, *AUC-Macro* sebesar 0,965, dan *AP-Macro* sebesar 0,887. Nilai tersebut menunjukkan bahwa representasi semantik yang dihasilkan oleh LSI mampu meningkatkan kemampuan model dalam membedakan berbagai kategori dokumen BKD secara lebih konsisten.

Model LSI + KNN menempati posisi kedua dengan nilai *Accuracy* sebesar 84,6%, *F1-Weighted* sebesar 0,848, *AUC-Macro* sebesar 0,930, dan *AP-Macro* sebesar 0,766. Walaupun masih menghasilkan performa yang baik, model ini menunjukkan sensitivitas terhadap distribusi data lokal sehingga performanya sedikit berada di bawah pendekatan berbasis *centroid.* Sementara itu, model TF-IDF + KNN memperoleh nilai *Accuracy* sebesar 82,6%, *F1-Weighted* sebesar 0,848, *AUC-Macro* sebesar 0,927, dan *AP-Macro* sebesar 0,783. Hasil tersebut menunjukkan bahwa penggunaan TF-IDF tanpa reduksi dimensi menyebabkan algoritma KNN bekerja pada ruang fitur yang sangat besar sehingga lebih rentan terhadap *noise* dan *fenomena curse of dimensionality.*

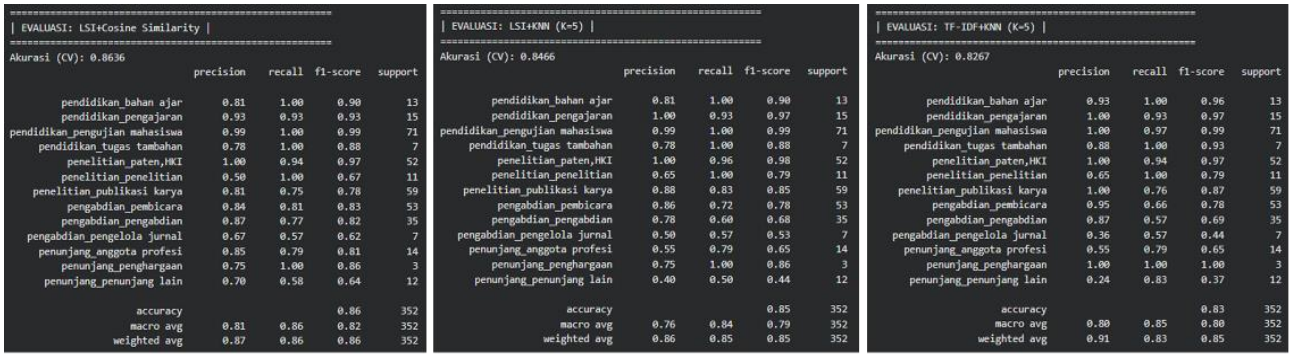
3.2 Implementasi/Pengujian

3.2.1 Hasil Pengujian Model

Model pertama yang diuji pada penelitian ini adalah kombinasi *Latent Semantic Indexing* (LSI) dengan *Cosine Similarity* berbasis *centroid*. Setelah seluruh dokumen direpresentasikan ke dalam ruang semantik menggunakan LSI, proses klasifikasi dilakukan dengan menghitung nilai *cosine similarity* antara dokumen uji terhadap *centroid* masing-masing kelas. Dokumen kemudian diklasifikasikan ke kelas yang memiliki nilai kemiripan tertinggi. Hasil evaluasi menunjukkan bahwa sebagian besar kategori utama berhasil diklasifikasikan dengan tingkat *precision* dan *recall* yang tinggi. Kategori Pendidikan dan Penelitian memiliki performa terbaik karena memiliki jumlah data yang relatif lebih banyak dibandingkan kategori lainnya. Sebaliknya, beberapa subkategori dengan jumlah data terbatas masih menunjukkan nilai *recall* yang lebih rendah. Model kedua menggunakan kombinasi representasi semantik LSI dengan algoritma *K-Nearest Neighbor.* Sebelum proses klasifikasi dilakukan, parameter terbaik ditentukan menggunakan *GridSearchCV* sehingga diperoleh nilai K = 5. Secara umum model mampu mengklasifikasikan sebagian besar dokumen dengan baik, namun performa pada kelas minoritas masih mengalami penurunan akibat karakteristik algoritma KNN yang sangat dipengaruhi oleh distribusi data di sekitar titik uji.

Sebagai pembanding, penelitian ini juga mengimplementasikan model *baseline* menggunakan representasi TF-IDF tanpa reduksi dimensi LSI. Pada model ini proses klasifikasi dilakukan langsung menggunakan algoritma KNN dengan nilai K = 5. Nilai akurasi tersebut merupakan yang terendah dibandingkan dua model lainnya. Hal ini menunjukkan bahwa representasi TF-IDF dengan dimensi fitur yang sangat tinggi belum mampu menggambarkan hubungan semantik antar dokumen secara optimal. Hasil pengujian dapat dilihat pada gambar 3.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274



| EVALUASI: LSI+Cosine Similarity |

Akurasi (CV): 0.8636

	precision	recall	f1-score	support
pendidikan_bahan ajar	0.81	1.00	0.90	13
pendidikan_pengajaran	0.93	0.93	0.93	15
pendidikan_pengujian mahasiswa	0.99	1.00	0.99	71
pendidikan_tugas tambahan	0.78	1.00	0.88	7
penelitian_paten,HKI	1.00	0.94	0.97	52
penelitian_penelitian	0.50	1.00	0.67	11
penelitian_publikasi karya	0.81	0.75	0.78	59
pengabdian_pembicara	0.84	0.81	0.83	53
pengabdian_pengabdian	0.87	0.77	0.82	35
pengabdian_pengelola jurnal	0.67	0.57	0.62	7
penunjang_anggota profesi	0.85	0.79	0.81	14
penunjang_penghargaan	0.75	1.00	0.86	3
penunjang_penunjang lain	0.70	0.58	0.64	12
accuracy			0.86	352
macro avg	0.81	0.86	0.82	352
weighted avg	0.87	0.86	0.86	352

| EVALUASI: LSI+KNN (K=5) |

Akurasi (CV): 0.8466

	precision	recall	f1-score	support
pendidikan_bahan ajar	0.81	1.00	0.90	13
pendidikan_pengajaran	1.00	0.93	0.97	15
pendidikan_pengujian mahasiswa	0.99	1.00	0.99	71
pendidikan_tugas tambahan	0.78	1.00	0.88	7
penelitian_paten,HKI	1.00	0.96	0.98	52
penelitian_penelitian	0.65	1.00	0.79	11
penelitian_publikasi karya	0.88	0.83	0.85	59
pengabdian_pembicara	0.86	0.72	0.78	53
pengabdian_pengabdian	0.78	0.60	0.68	35
pengabdian_pengelola jurnal	0.50	0.57	0.53	7
penunjang_anggota profesi	0.55	0.79	0.65	14
penunjang_penghargaan	0.75	1.00	0.86	3
penunjang_penunjang lain	0.40	0.50	0.44	12
accuracy			0.85	352
macro avg	0.76	0.84	0.79	352
weighted avg	0.86	0.85	0.85	352

| EVALUASI: TF-IDF+KNN (K=5) |

Akurasi (CV): 0.8267

	precision	recall	f1-score	support
pendidikan_bahan ajar	0.93	1.00	0.96	13
pendidikan_pengajaran	1.00	0.93	0.97	15
pendidikan_pengujian mahasiswa	1.00	0.97	0.99	71
pendidikan_tugas tambahan	0.88	1.00	0.93	7
penelitian_paten,HKI	1.00	0.94	0.97	52
penelitian_penelitian	0.65	1.00	0.79	11
penelitian_publikasi karya	1.00	0.76	0.87	59
pengabdian_pembicara	0.95	0.66	0.78	53
pengabdian_pengabdian	0.87	0.57	0.69	35
pengabdian_pengelola jurnal	0.36	0.57	0.44	7
penunjang_anggota profesi	0.55	0.79	0.65	14
penunjang_penghargaan	1.00	1.00	1.00	3
penunjang_penunjang lain	0.24	0.83	0.37	12
accuracy			0.83	352
macro avg	0.80	0.85	0.80	352
weighted avg	0.91	0.83	0.85	352

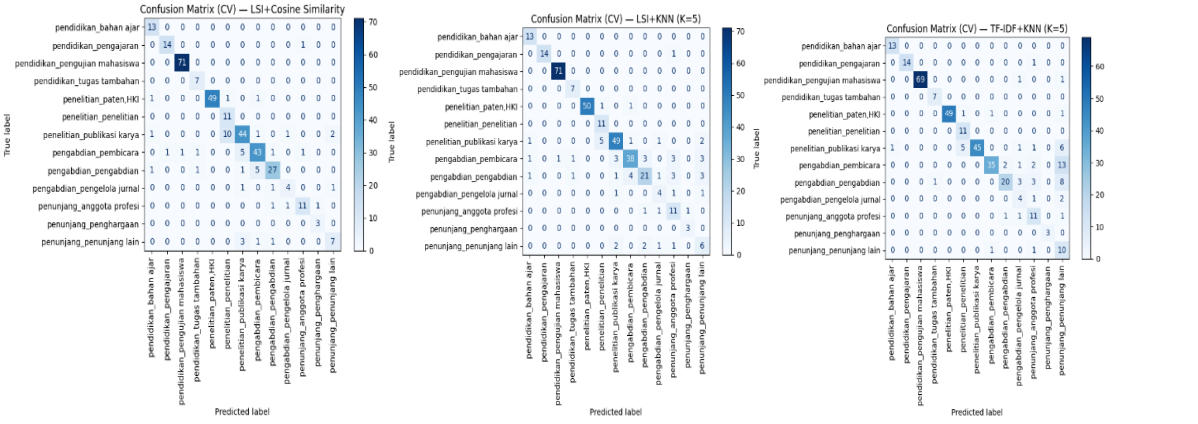
Gambar 3. *Classification Report*

3.2.2 *Confusion Matrix*

Untuk mengetahui pola prediksi setiap model secara lebih rinci, dilakukan analisis *menggunakan confusion matrix.* Hasil *confusion matrix* pada model LSI + *Cosine Similarity* menunjukkan bahwa sebagian besar prediksi berada pada diagonal utama. Kondisi tersebut menunjukkan bahwa sebagian besar dokumen berhasil diklasifikasikan ke kategori yang benar. Kesalahan klasifikasi relatif sedikit dan lebih banyak terjadi pada subkategori yang memiliki karakteristik semantik serupa.

Pada model LSI + KNN, pola diagonal utama masih terlihat dominan, namun jumlah kesalahan klasifikasi sedikit lebih besar dibandingkan model sebelumnya. Hal ini menunjukkan bahwa pendekatan berbasis tetangga terdekat lebih sensitif terhadap distribusi data lokal.

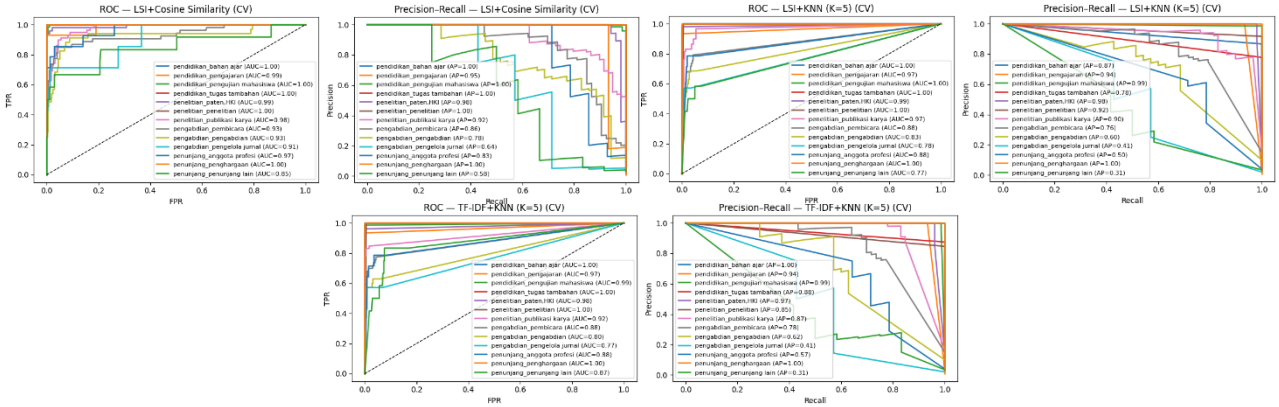
Sementara itu, *confusion matrix* model TF-IDF + KNN memperlihatkan penyebaran kesalahan prediksi yang lebih luas. Tanpa reduksi dimensi, representasi fitur menjadi sangat jarang (*sparse*) sehingga proses pencarian tetangga terdekat menjadi kurang optimal. Nilai *confusion matrix* dapat dilihat pada gambar 4.



Gambar 4. *Confusion Matrix*

3.2.3 Kurva ROC

Kurva *Receiver Operating Characteristic* (ROC) digunakan untuk mengukur kemampuan model dalam membedakan setiap kelas dokumen dapat dilihat pada gambar 5.



Gambar 5. Kurva ROC TF-IDF + KNN

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

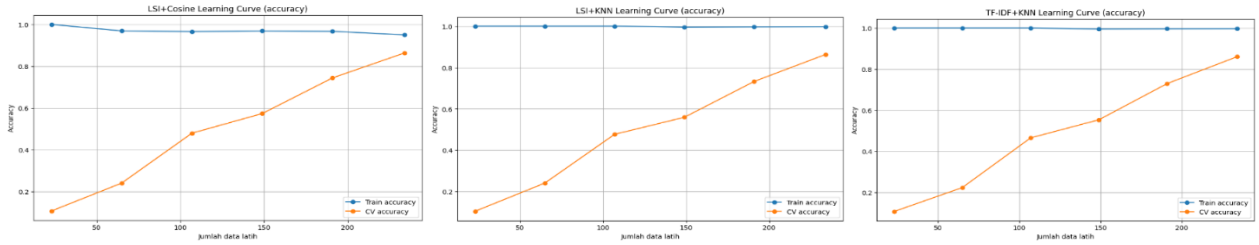
Hasil pengujian menunjukkan bahwa model LSI + *Cosine Similarity* memperoleh nilai *AUC-Macro* sebesar 0,965, yang menunjukkan kemampuan diskriminasi kategori yang sangat baik. Model LSI + KNN memperoleh nilai *AUC-Macro* sebesar 0,930, sedangkan model TF-IDF + KNN memperoleh nilai 0,927.

Seluruh model memiliki nilai AUC di atas 0,90 sehingga secara umum mampu membedakan kelas dokumen dengan baik, meskipun model LSI + *Cosine Similarity* memberikan performa terbaik.

3.2.4 *Learning Curve*

Learning curve digunakan untuk mengevaluasi kemampuan generalisasi model terhadap penambahan jumlah data pelatihan. Hasil pengujian menunjukkan bahwa model LSI + *Cosine Similarity* menghasilkan kurva pelatihan dan validasi yang semakin mendekat seiring bertambahnya jumlah data. Hal ini menunjukkan bahwa model memiliki kemampuan generalisasi yang baik. Pada model LSI + KNN, kurva pelatihan tetap berada pada tingkat akurasi yang sangat tinggi, sedangkan kurva validasi meningkat secara bertahap hingga mencapai nilai sekitar 85%. Kondisi tersebut menunjukkan bahwa performa model meningkat seiring bertambahnya jumlah data.

Model TF-IDF + KNN juga memperlihatkan peningkatan akurasi validasi ketika jumlah data bertambah, namun jarak antara kurva pelatihan dan validasi masih relatif besar sehingga menunjukkan kecenderungan *overfitting* yang lebih tinggi. Hasil nilai *learning curve* dapat dilihat pada Gambar 6.

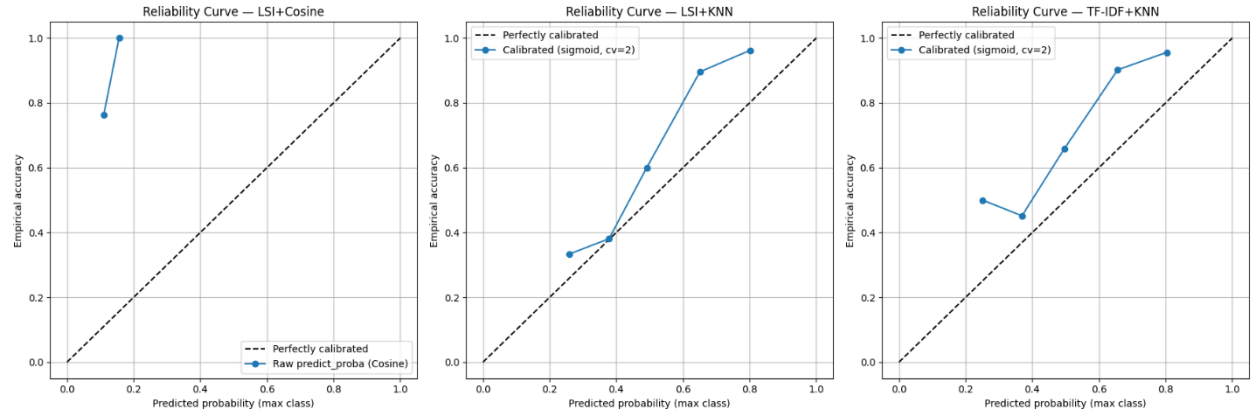


Gambar 6. *Learning Curve.*

3.2.5 *Reliability Curve*

Reliability curve digunakan untuk mengevaluasi kesesuaian antara probabilitas prediksi yang dihasilkan model dengan tingkat akurasi aktualnya. Hasil pengujian menunjukkan bahwa ketiga model memiliki karakteristik *underconfident,* yaitu probabilitas prediksi yang dihasilkan lebih rendah dibandingkan tingkat akurasi aktual.

Model LSI + *Cosine Similarity* menunjukkan pola kalibrasi yang paling stabil meskipun probabilitas prediksi berada pada rentang yang relatif rendah. Pada model LSI + KNN dan TF-IDF + KNN, proses *Sigmoid Calibration* (*Platt Scaling*) berhasil memperbaiki distribusi probabilitas sehingga lebih mendekati kondisi ideal. Hasil nilai *reliability curve* dapat dilihat pada Gambar 7.



Gambar 7. *Reliability Curve*

Hasil pengujian menunjukkan bahwa pendekatan *Latent Semantic Indexing* (LSI) yang dikombinasikan dengan *Cosine Similarity* memberikan performa terbaik dibandingkan model LSI + *K-Nearest Neighbor* (KNN) maupun model baseline TF-IDF + KNN. Model LSI + *Cosine Similarity* memperoleh nilai *accuracy* sebesar 86,3%, *F1-Weighted* sebesar 0,865, *AUC-Macro* sebesar 0,965, dan *Average Precision* (*AP-Macro*) sebesar 0,887. Sementara itu, model LSI + KNN memperoleh *Accuracy* sebesar 84,6%, *F1-Weighted* sebesar 0,848, *AUC-Macro* sebesar 0,930, dan *AP-Macro* sebesar 0,766 dan model TF-IDF + KNN menghasilkan *Accuracy* sebesar 82,6%, *F1-Weighted* sebesar 0,848, *AUC-Macro* sebesar 0,927, dan *AP-Macro* sebesar 0,783.

Perbedaan performa tersebut menunjukkan bahwa penggunaan representasi semantik melalui LSI memberikan kontribusi yang signifikan dalam meningkatkan kualitas klasifikasi dokumen BKD. Berbeda dengan TF-IDF yang hanya merepresentasikan frekuensi kemunculan kata, LSI mampu menangkap hubungan semantik antar istilah sehingga dokumen yang menggunakan kosakata berbeda tetapi memiliki makna yang sama tetap dapat direpresentasikan secara

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1988-2000
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1274

berdekatan dalam ruang fitur. Kondisi ini sangat sesuai dengan karakteristik dokumen BKD yang berasal dari berbagai jenis aktivitas tridharma sehingga memiliki variasi istilah yang cukup tinggi.

Keunggulan lain terlihat pada penggunaan *Cosine Similarity* berbasis *centroid*. Pendekatan ini menentukan kategori dokumen berdasarkan tingkat kemiripan antara dokumen uji dengan *centroid* masing-masing kelas. Karena *centroid* dibentuk dari rata-rata representasi seluruh dokumen pada setiap kategori, proses klasifikasi menjadi lebih stabil terhadap variasi dokumen individu. Pendekatan tersebut juga lebih tahan terhadap ketidakseimbangan distribusi data dibandingkan algoritma berbasis tetangga terdekat.

Sebaliknya, model LSI + KNN masih menunjukkan penurunan performa pada beberapa kategori minoritas. Walaupun representasi semantik telah diperbaiki melalui LSI, proses pengambilan keputusan KNN tetap bergantung pada kedekatan sejumlah tetangga terdekat. Pada kondisi dataset yang tidak seimbang, dokumen dari kelas mayoritas cenderung mendominasi lingkungan sekitar data uji sehingga peluang salah klasifikasi terhadap kelas minoritas menjadi lebih besar. Oleh karena itu, teknik SMOTE diterapkan untuk menghasilkan distribusi data latih yang lebih seimbang pada model KNN.

Model *baseline* TF-IDF + KNN memperoleh performa paling rendah. Kondisi ini menunjukkan bahwa penggunaan representasi TF-IDF secara langsung masih menghadapi permasalahan tingginya dimensi fitur (*high dimensionality)* dan *sparsity*. Dalam ruang fitur berdimensi tinggi, jarak antar dokumen menjadi kurang representatif sehingga proses pencarian tetangga terdekat tidak mampu menggambarkan kemiripan semantik secara optimal. Akibatnya, model lebih banyak bergantung pada kecocokan kata secara literal dibandingkan kesamaan makna antar dokumen.

Hasil *confusion matrix* juga memperlihatkan bahwa sebagian besar kesalahan klasifikasi terjadi pada kategori dengan jumlah data yang relatif sedikit, seperti Pengabdian–Pengelola Jurnal dan Penunjang–Penghargaan. Hal ini menunjukkan bahwa jumlah data latih masih menjadi faktor penting dalam meningkatkan kemampuan model mengenali karakteristik setiap kategori. Sebaliknya, kategori yang memiliki jumlah dokumen lebih banyak, seperti Pendidikan–Pengujian Mahasiswa dan Penelitian–Publikasi Karya, berhasil diklasifikasikan dengan tingkat akurasi yang jauh lebih tinggi.

Analisis kurva ROC menunjukkan bahwa seluruh model memiliki kemampuan diskriminasi kategori yang baik dengan nilai *AUC-Macro* di atas 0,90. Namun demikian, model LSI + *Cosine Similarity* tetap memperoleh nilai tertinggi sebesar 0,965. Nilai tersebut mengindikasikan bahwa model mampu membedakan setiap kategori dokumen dengan tingkat sensitivitas yang sangat baik.

Hasil *learning curve* menunjukkan bahwa bertambahnya jumlah data pelatihan memberikan peningkatan performa validasi pada seluruh model. Pada model LSI + *Cosine Similarity*, jarak antara kurva pelatihan dan validasi relatif kecil sehingga mengindikasikan kemampuan generalisasi yang baik. Sebaliknya, pada model KNN, akurasi pelatihan cenderung mendekati 100% sementara akurasi validasi lebih rendah, yang menunjukkan adanya kecenderungan *overfitting* ringan.

Analisis *reliability curve* juga memperlihatkan bahwa ketiga model memiliki karakteristik *underconfident*, yaitu probabilitas prediksi yang dihasilkan lebih rendah dibandingkan tingkat akurasi aktual. Meskipun demikian, setelah dilakukan kalibrasi menggunakan *Sigmoid Calibration* pada model KNN, distribusi probabilitas menjadi lebih representatif terhadap kondisi sebenarnya.

3.3 Pembahasan

Hasil penelitian ini menunjukkan bahwa penerapan *Latent Semantic Indexing* (LSI) mampu meningkatkan performa klasifikasi dokumen BKD dibandingkan penggunaan representasi TF-IDF secara langsung. Hal tersebut ditunjukkan oleh model LSI + *Cosine Similarity* yang memperoleh nilai akurasi sebesar 86,3%, lebih tinggi dibandingkan LSI + KNN sebesar 84,6% dan model *baseline* TF-IDF + KNN sebesar 82,6%. Hasil tersebut mengindikasikan bahwa proses reduksi dimensi menggunakan LSI mampu menghasilkan representasi dokumen yang lebih informatif sehingga mempermudah proses klasifikasi.

Temuan penelitian ini sejalan dengan penelitian Istighfarizky et al. (2022) [3] yang menerapkan kombinasi LSI dan KNN pada klasifikasi dokumen abstrak ilmiah berbahasa Indonesia dan memperoleh akurasi sebesar 84,3%. Pada penelitian tersebut, LSI berperan dalam mereduksi dimensi ruang fitur sehingga algoritma KNN dapat bekerja pada representasi data yang lebih ringkas. Hasil penelitian ini menunjukkan kecenderungan yang sama, yaitu penggunaan LSI mampu menghasilkan performa klasifikasi yang lebih baik dibandingkan apabila KNN bekerja langsung pada ruang fitur TF-IDF berdimensi tinggi.

Hasil penelitian ini juga konsisten dengan penelitian Murniati (2021) [4] yang mengombinasikan LSI dengan *Cosine Similarity* pada analisis sentimen aplikasi *mobile banking* dan memperoleh akurasi sebesar 81,67%. Pada penelitian tersebut, *Cosine Similarity* digunakan untuk mengukur kedekatan antar dokumen dalam ruang semantik hasil reduksi dimensi LSI. Pada penelitian ini, pendekatan yang sama diterapkan pada dokumen BKD dan menghasilkan performa yang lebih tinggi dengan akurasi 86,3%, sehingga menunjukkan bahwa kombinasi LSI dan *Cosine Similarity* juga efektif diterapkan pada dokumen administrasi akademik.

Penelitian Rivaldi et al. (2024) [5] juga menyatakan bahwa kombinasi LSI dan KNN mampu meningkatkan stabilitas klasifikasi karena reduksi dimensi dapat mengurangi *noise* yang memengaruhi proses pencarian tetangga terdekat. Temuan tersebut sejalan dengan hasil penelitian ini, di mana model LSI + KNN menghasilkan performa yang lebih baik dibandingkan model TF-IDF + KNN, meskipun masih berada di bawah pendekatan LSI + *Cosine Similarity.*

Selanjutnya, penelitian Ajijah dan Kurniawan (2023) [6] menunjukkan bahwa metode LSI-*Cosine Similarity* efektif digunakan untuk mengukur kemiripan semantik pada dokumen administrasi pendidikan tinggi. Hasil tersebut mendukung temuan penelitian ini, karena dokumen BKD juga merupakan dokumen administrasi akademik yang memiliki banyak istilah teknis dan variasi kosakata. Penggunaan LSI mampu menangkap hubungan semantik antar istilah sehingga pendekatan *Cosine Similarity* dapat memberikan hasil klasifikasi yang lebih akurat.

Sementara itu, penelitian Kosasih dan Alberto (2021) [7] yang menggunakan TF-IDF dan KNN memperoleh akurasi sebesar 79,33%, sedangkan penelitian Mehta et al. (2023) [8] juga menunjukkan bahwa algoritma KNN memiliki kemampuan yang baik dalam proses klasifikasi dokumen teks. Namun demikian, hasil penelitian ini menunjukkan bahwa performa KNN masih dipengaruhi oleh representasi fitur yang digunakan. Ketika KNN diterapkan langsung pada ruang fitur TF-IDF, performanya lebih rendah dibandingkan ketika menggunakan representasi hasil reduksi dimensi LSI.

Perbedaan utama penelitian ini dibandingkan penelitian-penelitian sebelumnya terletak pada objek penelitian dan pendekatan yang digunakan. Penelitian terdahulu umumnya menggunakan dokumen berupa abstrak ilmiah, analisis sentimen, maupun dokumen teks umum, sedangkan penelitian ini menggunakan dokumen Beban Kerja Dosen (BKD) yang memiliki karakteristik khusus berupa banyaknya subkategori tridharma, variasi istilah akademik, serta distribusi data yang tidak seimbang. Selain itu, penelitian ini tidak hanya mengevaluasi satu metode, tetapi membandingkan secara langsung tiga pendekatan, yaitu LSI + *Cosine Similarity*, LSI + KNN, dan TF-IDF + KNN sebagai model pembanding (*baseline*). Penelitian ini juga menambahkan proses penambahan kata kunci (*keyword enrichment*) pada tahap pra-pemrosesan untuk memperkuat representasi dokumen sebelum dilakukan pembobotan TF-IDF dan reduksi dimensi menggunakan LSI. Dengan demikian, penelitian ini memberikan kontribusi baru dalam penerapan metode klasifikasi pada dokumen BKD di lingkungan Pascasarjana Universitas Negeri Medan.

Penelitian ini memiliki beberapa kelebihan. Pertama, penelitian tidak hanya membandingkan dua algoritma klasifikasi, tetapi juga mengevaluasi pengaruh representasi fitur terhadap performa model melalui perbandingan antara TF-IDF dan LSI. Kedua, proses pra-pemrosesan dilakukan secara lengkap mulai dari ekstraksi teks, normalisasi, *tokenisasi, stopword removal, stemming*, hingga penambahan kata kunci berbasis aturan yang mampu memperkuat representasi dokumen BKD. Ketiga, hasil penelitian telah diimplementasikan ke dalam sistem berbasis web sehingga model dapat digunakan secara langsung dalam proses klasifikasi dokumen BKD.

Di samping kelebihannya, penelitian ini juga memiliki beberapa keterbatasan. Dataset yang digunakan masih berasal dari satu institusi, yaitu Pascasarjana Universitas Negeri Medan, sehingga variasi dokumen belum sepenuhnya merepresentasikan kondisi pada perguruan tinggi lain. Selain itu, distribusi data antar kategori masih belum seimbang meskipun telah dilakukan penanganan menggunakan SMOTE pada model KNN. Jumlah dokumen pada beberapa subkategori masih relatif sedikit sehingga memengaruhi kemampuan model dalam mengenali kelas minoritas. Keterbatasan lainnya adalah penggunaan metode klasifikasi yang masih berbasis representasi vektor klasik. Penelitian selanjutnya dapat membandingkan performa LSI dengan pendekatan representasi modern berbasis *word embedding* maupun *transformer*, seperti *FastText, Doc2Vec, BERT,* atau *IndoBERT,* untuk mengetahui peningkatan performa yang dapat diperoleh pada klasifikasi dokumen BKD.

4. KESIMPULAN

Penelitian ini berhasil mengevaluasi dan membandingkan kinerja tiga pendekatan klasifikasi dokumen Beban Kerja Dosen (BKD), yaitu *Latent Semantic Indexing* (LSI) + *Cosine Similarity*, LSI + *K-Nearest Neighbor* (KNN), dan TF-IDF + KNN sebagai *baseline*, pada dataset 352 dokumen BKD Pascasarjana Universitas Negeri Medan yang telah melalui tahapan ekstraksi teks, pra-pemrosesan, pembobotan TF-IDF, serta reduksi dimensi menggunakan LSI. Hasil pengujian menunjukkan bahwa pendekatan LSI + *Cosine Similarity* memberikan performa terbaik dengan *Accuracy* 86,3%, *F1-Weighted* 0,865, *AUC-Macro* 0,965, dan *AP-Macro* 0,887, diikuti oleh LSI + KNN dengan *Accuracy* 84,6%, dan TF-IDF + KNN sebagai *baseline* dengan *Accuracy* 82,6%. Hasil tersebut menjawab tujuan penelitian bahwa representasi semantik menggunakan LSI mampu meningkatkan kualitas klasifikasi dokumen BKD dibandingkan representasi TF-IDF konvesional, serta menunjukkani bahwa pendekatan berbasis *Cosine Similarity* lebih stabil dibandingkan KNN karena tidak terlalu dipengaruhi oleh distribusi lokal data maupun ketidakseimbangan jumlah dokumen pada setiap kelas. Meskipun demikian, penelitian ini masih memiliki keterbatasan karena dataset hanya berasal dari satu institusi dan distribusi data antar kategori masih belum sepenuhnya seimbang sehingga performa pada beberapa kelas minoritas relatif lebih rendah. Oleh karena itu, penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan beragam, menerapkan Teknik penanganan ketidakseimbangan data yang lebih optimal, serta membandingkan pendekatan LSI dengan metode representasi modern berbasis *word embedding* maupun *transformer* seperti *Word2Vec, FastText, BERT,* atau *IndoBERT* untuk memperoleh model klasifikasi dokumen BKD yang lebih akurat, general, dan tahan terhadap variasi karakteristik dokumen akademik.

REFERENCES

[1] M. Solekhah and W. Lasniah, "Analisis Proses Bisnis Sistem Informasi Manajemen Dokumen Pendukung Beban Kerja Dosen Dengan Metode Protoyping Model," *Sebatik*, vol. 25, no. 2, pp. 356–365, 2021, doi: 10.46984/sebatik.v25i2.1656.

[2] A. H. Ardiansyah, K. P. Kartika, and S. N. Budiman, "Penerapan Latent Semantic Indexing Pada Sistem Temu Balik Informasi Pada Undang-Undang Pemilu Berdasarkan Kasus," *J. Mnemon.*, vol. 4, no. 2, pp. 64–70, 2021, doi:

10.36040/mnemonic.v4i2.4165.

[3] F. Istighfarizky, N. A. Sanjaya ER, I. M. Widiartha, L. G. Astuti, I. G. N. A. C. Putra, and I. K. G. Suhartana, "Klasifikasi Jurnal menggunakan Metode KNN dengan Mengimplementasikan Perbandingan Seleksi Fitur," *Jeliku (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 11, no. 1, p. 167, 2022, doi: 10.24843/jlk.2022.v11.i01.p18.

[4] R. A. Murniati, "Analisis Sentimen Ulasan Aplikasi Mobile Banking M-Smile Menggunakan Metode Latent Semantic Indexing (LSI)," Universitas Paramadina, 2021. [Online]. Available: https://repository.paramadina.ac.id/1483/1/Jurnal_Retno Asih M_120103007.pdf

[5] R. C. Rivaldi and T. D. Wismarini, "Analisis Sentimen Pada Ulasan Produk Dengan Metode Natural Language Processing (NLP) (Studi Kasus Zalika Store 88 Shopee)," *Elkom J. Elektron. dan Komput.*, vol. 17, no. 1, pp. 120–128, 2024, doi: 10.51903/elkom.v17i1.1680.

[6] N. Ajijah, A. Kurniawan, and Susilawati, "Klasifikasi Teks Mining Terhadap Analisa Isu Kegiatan Tenaga Lapangan Menggunakan Algoritma K-Nearest Neighbor (KNN)," *J-Sakti (Jurnal Sains Komput. Inform.*, vol. 7, no. 1, pp. 254–262, 2023, doi: 10.30645/j-sakti.v7i1.589.

[7] R. Kosasih and A. Alberto, "Analisis Sentimen Produk Permainan Menggunakan Metode TF-IDF Dan Algoritma K-Nearest Neighbor," *InfoTekJar J. Nas. Inform. dan Teknol. Jar.*, vol. 6, no. 1, pp. 134–139, 2021, doi: 10.30743/infotekjar.v6i1.3893.

[8] P. Mehta, S. Aggarwal, and A. Tandon, "The Effect of Topic Modelling on Prediction of Criticality Levels of Software Vulnerabilities," *Inform.*, vol. 47, no. 6, pp. 145–158, 2023, doi: 10.31449/inf.v47i6.3712.

[9] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *Klik Kaji. Ilm. Inform. Dan Komput.*, vol. 2, no. 5, pp. 178–185, 2022, doi: 10.30865/klik.v2i5.362.

[10] O. W. Yuda, D. Tuti, L. S. Yee, and Susanti, "Penerapan Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest," *Satin - Sains dan Teknol. Inf.*, vol. 8, no. 2, pp. 122–131, 2022, doi: 10.33372/stn.v8i2.885.

[11] A. Dahari, D. Herwanto, and J. Arifin, "Analisa Pengendalian Persediaan Bahan Baku Bumbu Racik Makanan dari Raw Material Hingga Barang Jadi (Finish Good) di PT. Ariake Europe Indonesia," *J. Ilm. Wahana Pendidik.*, vol. 7, no. 1, pp. 391–402, 2021, doi: 10.5281/zenodo.5535550.

[12] J. Jefriyanto, N. Ainun, and M. A. Al Ardha, "Application of Naïve Bayes Classification to Analyze Performance Using Stopwords," *Jiste (Journal Inf. Syst. Technol. Eng.*, vol. 1, no. 1, pp. 49–53, 2023, doi: 10.61487/jiste.v1i2.15.

[13] A. Sinaga and S. P. Nainggolan, "Analisis Perbandingan Akurasi Dan Waktu Proses Algoritma Stemming Arifin-Setiono Dan Nazief-Adriani Pada Dokumen Teks Bahasa Indonesia," *Sebatik*, vol. 27, no. 1, pp. 63–69, 2023, doi: 10.46984/sebatik.v27i1.2072.

[14] A. R. Lubis and M. K. M. Nasution, "Twitter Data Analysis and Text Normalization in Collecting Standard Word," *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 2, pp. 855–863, 2023, doi: 10.37385/jaets.v4i2.1991.

[15] Supiyanto and Sriyono, "Metode Cosine Similarity Untuk Mendeteksi Kemiripan Pada Dokumen Teks," *J. Mipa dan Pengajarannya*, vol. 1, no. 1, pp. 1–7, 2023, doi: 10.31957/sains.v23i1.3661.

[16] R. F. Putra *et al.*, *Algoritma Pembelajaran Mesin (Dasar , Teknik , dan Aplikasi)*, Pertama., no. April. Bekasi: PT. Sonpedia Publishing Indonesia, 2024. [Online]. Available: https://www.researchgate.net/publication/379479664_ALGORITMA_PEMBELAJARAN_MESIN_Dasar_Teknik_dan_Aplikasi

[17] A. Muzakir, A. Desiani, and A. Amran, "Klasifikasi Penyakit Kanker Prostat Menggunakan Algoritma Naïve Bayes Classification of Prostate Cancer Using Naïve Bayes and K-Nearest Neighbor Algorithms," *Komputika J. Sist. Komput.*, vol. 12, no. 148, pp. 73–79, 2023, doi: 10.34010/komputika.v12i1.9629.

[18] W. P. Anggraini, M. S. Utami, J. M. Berlianty, E. S. Hutagalung, Y. Juniarto, and R. Nooraeni, "Klasifikasi Sentimen Masyarakat Terhadap Kebijakan Kartu Pekerja Di Indonesia," *Fakt. Exacta*, vol. 13, no. 4, p. 255, 2021, doi: 10.30998/faktorexacta.v13i4.7964.

[19] R. R. Adhitya, W. Witanti, and R. Yuniarti, "Perbandingan Metode Cart Dan Naïve Bayes Untuk Klasifikasi Customer Churn," *Infotech J.*, vol. 9, no. 2, pp. 307–318, 2023, doi: 10.31949/infotech.v9i2.5641.

[20] M. C. Rani, F. D. Azkia, R. A. Dewi, M. Wahyudi, Sumanto, and A. S. Budiman, "Perbandingan Algoritma Random Forest, Naive Bayes, dan Neural Network dalam Klasifikasi Penyakit Jantung," *J. Sains Inform. Terap. (Jsit)*, vol. 4, no. 2, pp. 187–201, 2025, doi: 10.62357/jsit.v4i2.609.

[21] Angelina, N. Filosofia, and R. Arga Wijaya, "Optimizing Predictions For Thyroid Disease Sufferers Using Correlation Matrix And Random Forest With Hyperparameter Tuning," *Media J. Gen. Comput. Sci.*, vol. 2, no. 1, pp. 1–11, 2025, doi: 10.62205/mjgcs.v2i1.26.

[22] W. A. Kurniawan and A. Salam, "Penggunaan Feature Space Smote Untuk Mengurangi Overfitting Akibat Imbalance Dataset," *Techno.Com*, vol. 23, no. 2, pp. 328–337, 2024, doi: 10.62411/tc.v23i2.10215.

[23] I. W. Amitaba, A. Pramudiya, H. Ahyadi, and L. S. C. Ranesa, "Optimasi Pemodelan Hujan Limpasan dan Analisis Genangan Banjir untuk Evaluasi Dampak Infrastruktur DAS Rea," *J. Sumber Daya Air*, vol. 21, no. 2, pp. 101–114, 2025, doi: 10.32679/jsda.v21i2.970.