Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499

Implementasi LDA, TF-IDF, dan BERT dalam Sistem Rekomendasi Dosen Pembimbing untuk Mahasiswa

Mutiara Syabilla*, Junta Zeniarja, Qotrunnada Nabila

Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: 1,*mutiarasyabilla13@gmail.com, 2junta@dsn.dinus.ac.id, 3nadanabila2172@gmail.com
Email Penulis Korespondensi: mutiarasyabilla13@gmail.com
Submitted: **22/12/2024**; Accepted: **26/02/2025**; Published**: 01/03/2025**

Abstrak−Pemilihan dosen pembimbing sering kali dilakukan secara manual, akibatnya sering memakan waktu dalam mencocokkan topik penelitian mahasiswa dengan keahlian dosen. Penelitian ini mengembangkan sistem rekomendasi dosen pembimbing berbasis judul dan abstrak tugas akhir mahasiswa yang mengintegrasikan *Latent Dirichlet Allocation (LDA)*, *Term Frequency-Inverse Document Frequency (TF-IDF)*, dan *Bidirectional Encoder Representations from Transformers (BERT)*. Dataset penelitian mencakup 1.096 data dari 71 dosen Teknik Informatika Universitas Dian Nuswantoro, yang dikumpulkan melalui Google Scholar. Tahapan analisis dimulai dengan pemrosesan teks seperti *case folding*, tokenisasi, dan *stemming*, diikuti dengan analisis topik menggunakan LDA, pemberian bobot kata spesifik melalui TF-IDF, dan representasi vektor kaya konteks menggunakan BERT. Model ini mencocokkan topik penelitian mahasiswa dengan keahlian dosen menggunakan *Cosine Similarity*. Evaluasi menunjukkan tingkat akurasi sebesar 80%, *presisi* 66%, dan *recall* 19%, yang menggambarkan bahwa model mampu memberikan rekomendasi yang akurat meskipun masih terdapat item relevan yang terlewatkan. Model ini memberikan hasil yang efektif dalam memfasilitasii pemilihan dosen pembimbing. Penelitian ini diharapkan membantu mahasiswa dalam menemukan dosen pembimbing serta mempermudah dosen dalam mengidentifikasi mahasiswa dengan minat penelitian yang relevan.

Kata Kunci: Rekomendasi pembimbing; LDA; TF-IDF; BERT; Cosine Similarity

Abstract−The selection of thesis supervisors is often done manually, which tends to be time-consuming in matching students' research topics with the expertise of faculty members. This study develops a thesis supervisor recommendation system based on the title and abstract of students' final projects, integrating Latent Dirichlet Allocation (LDA), Term Frequency-Inverse Document Frequency (TF-IDF), and Bidirectional Encoder Representations from Transformers (BERT). The research dataset includes 1,096 records from 71 faculty members in the Informatics Engineering Department at Universitas Dian Nuswantoro, collected through Google Scholar. The analysis process begins with text preprocessing such as case folding, tokenization, and stemming, followed by topic analysis using LDA, term-specific weighting through TF-IDF, and context-rich vector representation using BERT. The model matches students' research topics with faculty expertise using Cosine Similarity. Evaluation results show an accuracy of 80%, precision of 66%, and recall of 19%, indicating that the model can provide accurate recommendations, though some relevant items are still missed. This model proves effective in facilitating the selection of thesis supervisors. This research is expected to assist students in finding suitable supervisors and help faculty members identify students with relevant research interests.

Keywords: Supervisor recommendation; LDA; TF-IDF; BERT; Cosine Similarity

1. PENDAHULUAN

Skripsi atau tugas akhir merupakan hasil penelitian yang dibuat oleh mahasiswa mengenai suatu permasalah sesuai bidang keilmuan dari mahasiswa, dengan bimbingan dari dosen pembimbing. Penulisan tugas akhir ini menjadi syarat wajib untuk menyelesaikan pendidikan dan memperoleh gelara ademik di perguruan tinggi [1]. Salah satu tahap penting dalam menyelesaikan tugas akhir adalah pemilihan dosen pembimbing yang tepat. Karena bimbingan dari dosen pembimbing yang tepat menjadi faktor penentu keberhasilan mahasiswa dalam menyelesaikan tugas akhir dengan baik [2][3]. Namun, banyak mahasiswa sering menghadapi tantangan dalam menemukan dosen pembimbing yang relevan dengan topik tugas akhir mereka. Proses ini seringkali dilakukan secara manual, yang berarti mahasiswa harus mencari sendiri informasi tentang keahlian dosen yang relevan dengan topik tugas akhir mereka [2]. Hal ini menjadi kendala karena tidak semua mahasiswa memiliki akses atau pengetahuan yang cukup untuk melakukan pencarian tersebut secara efektif. Keterbatasan ini mengakibatkan mahasiswa mengambil keputusan yang kurang informatif, seperti memilih dosen yang tidak memiliki pengalaman atau ketertarikan dalam bidang penelitian yang diinginkan. Selain itu tanpa adanya referensi yang jelas untuk mencocokkan topik penelitian dengan keahlian dosen proses menjadu tidak efisien dan mengakibatkan kombinasi yang tidak optimal antara mahasiswa dan dosen pembimbing [4].

Untuk mengatasi masalah ini, diperlukan sebuah model rekomendasi dosen pembimbing secara otomatis. Pemilihan dosen yang tepat akan meningkatkan kualitas bimbingan dan membantu mahasiswa menyelesaikan penelitian dengan lebih baik [5]. Dengan memanfaatkan teknologi terkirini, seperti teknik pemrosesan bahasa alami diharapkan mampu mencocokkan topik tugas akhir mahasiswa dengan keahlian dosen secara otomatis dan efisien [6]. Teknologi pemrosesan bahasa alami *(Natural Language Processing)* menunjukkan potensi untuk mengatasi masalah serupa di berbagai bidang. Misalkan, penelitian sebelumnya telah menunjukkan bahwa penggabungan model vektorisasi TF-IDF *(Term Frequency-Inverse Document Frequency)* dengan BERT *(Bidirectional Encoder Representations from Transformer)* mampu mencapai performa terbaik dalam aplikasi analisis teks [7]. Penelitian sebelumnya oleh Sun et al telah mengeksplorasi penggunaan BERT dan TF-IDF. Hasil penelitian menunjukkan bahwa metode BERT dan TF-IDF mengungguli model lain yaitu LSTM, CNN, dan BERT dengan peningkatan signifikan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499

dalam metrik klasifikasi [8]. Dalam konteks yang berbeda, metode semi-supervised juga telah dimanfaatkan oleh Yang et al untuk mengembangkan sistem rekomendasi literatur ilmiah yang berbasis BERT dan LDA. Penelitian ini berhasil menunjukkan bahwa BERT mampu menangkap konteks dalam dokumen, sedangkan LDA berperan dalam ekstraksi topik [9]. Penelitian oleh Jin et al juga melakukan penegembangan analisis sentimen menggunakan penggabungan BERT dan TF-IDF yang digunakan untuk klasifikasi multi-label. Penelitian menunjukkan bahwa metode yang diusulkan lebih efektif dibandingkan dengan metode tradisional [10].

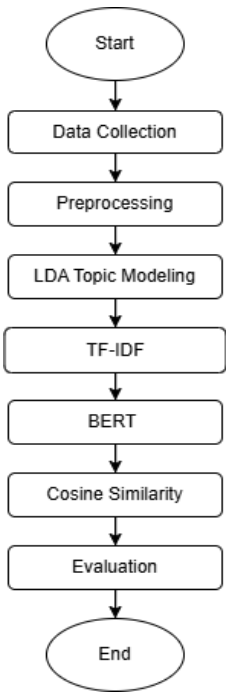
Sementara itu, Penelitian oleh Pérez et al melakukan pendekatan menggunakan LDA dan embedding BERT yang diterapkan untuk klasifikasi topik berita pariwisata yang tidak berlabel [11]. Pada rekomendasi berbasis ulasan, penelitian yang dilakukan oleh Zhang et al yang menggunakan kombinasi BERT dan LDA untuk menganalisis ulasan film di platform Douban menunjukkan hasil yang menjanjikan. Sistem ini memanfaatkan BERT untuk analisis sentimen dari ulasan, sedangkan LDA digunakan untuk ekstraksi topik dari kategori-kategori yang relevan. Model ini menghasilkan akurasi prediksi sentimen yang cukup baik[12]. Kombinasi model LDA dengan BERT telah menghasilkan hasil kinerja yang baik, terutama dalam mengukur kemiripan menggunakan cosine similarity [13].

Dalam pengembangan rekomendasi dosen pembimbing, pemanfaatan teknologi NLP dapat digunakan. Buku “Specch and Languange Processing” oleh Daniel Jurafsky dan James H. Martin memberikan landasan teori untuk pemahaman mengenai pemrosesan bahasa alami, termasuk teknik TF-IDF dan LDA [14]. Selain itu, “Transformers for Natural Language Processing” oleh Denis Rothman menyediakan pengetahuan mengenai model transformer seperti BERT [15]. Pemahaman ini relevan dengan konteks pengembangan rekomendasi yang mencocokan topik penelitian mahasiswa dengan topik yang diampu oleh dosen pembimbing.

Penelitian ini bertujuan untuk mengembangkan model rekomendasi dosen pembimbing yang mengintegrasikan ekstraksi topik menggunakan LDA dengan penggabungan vektorisasi TF-IDF dan embedding BERT, serta mengukur kemiripan antar dokumen dengan menggunakan *Cosine Similarity*. Dalam model ini, teknik analisis teks akan diterapkan untuk menganalisis judul dan abstrak penelitian mahasiswa sehingga dapat dihasilkan rekomendasi dosen yang relevan. Penelitian diharapkan mampu membantu tidak hanya mahasiswa dalam menemukan pembimbing yang tepat tetapi juga bagi dosen dalam mengidentifikasi potensi penelitian yang relevan dengan keahlian mereka [5]. Dengan demikian, penelitian ini berkontribusi pada pengembangan pendidikan tinggi yang lebih baik.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan metode *text similarity* dengan memanfaatkan data riwayat penelitian dosen pembimbing. Proses penelitian dimulai dengan pengumpulan dan pemrosesan data menggunakan *Natural Language Processing* (NLP) untuk menghasilkan klasifikasi dosen pembimbing. Diagram alur yang menunjukkan langkah-langkah penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Alur penelitian

2.1 Data Collection

Pengumpulan data dilakukan dengan metode scrapping, yaitu mengambil data penelitian dosen dari Google Scholar menggunaka bahasa Python di VSCode [16]. Pada proses scrapping, *library* Python seperti BeautifulSoup digunakan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499

untuk memparsing HTML di halaman profil Google Scholar. *Library* request untuk melakukan permintaan HTTP dengan memanfaatkan nama dosen pembimbing di Proses pengumpulan data dilakukan dengan memanfaatkan daftar nama-nama dosen yang menjadi dosen pembimbing di Teknik Informatika Universitas Dian Nuswantoro sebagai kata kunci pencarian [17]. Data yang dikumpulkan berupa nama, judul, abstrak, tahun, dan nama penulis lainnya. Hanya publikasi dalam lima tahun terakhir yaitu 2019-2024 yang diambil, dengan syarat nama dosen menjadi penulis pertama atau kedua. Hal ini dilakukan untuk memastikan bahwa penelitian yang dikumpulkan relevan dan terkini, serta mencerminkan keterlibatan aktif dosen dalam penelitian. Setelah data terkumpul, dilakukan pengecekan secara manual untuk memastikan bahwa judul dan abstrak sesuai dengan keahlian dosen. Selain itu, penelitian ini mematuhi kebijakan platform Google Scholar dengan hanya mengakses data publik saja.

2.2 Preprocessing

Preprocessing merupakan langkah penting dalam pemrosesan bahasa alami untuk membersihkan dan mempersiapkan data, tujuan untuk memudahkan analisis data [18]. Pada teks berbahasa indonesia, langkah yang dilakukan yaitu *case folding* untuk mengubah semua huruf menjadi huruf kecil. Selain itu, *punctuation removal* dilakukan untuk menghapus karakter yang tidak diperlukan, seperti angka, tanda baca, atau simbol, sehingga teks hanya berisi huruf dan spasi. Penghapusan *stopword* atau kata-kata umum seperti "dan", "untuk" dihapus menggunakan *library* Sastrawi, karena kata tersebut tidak memiliki makna dalam analisis. Langkah penting yaitu *stemming* untuk menghilangkan imbuhan awalan dan akhiran, sehingga kata kembali ke bentuk dasar. Misalkan, kata "penelitian" diubah menjadi "teliti" menggunakan *library* Sastrawi[19][20]. Sementara itu untuk teks berbahasa inggris memiliki langkah serupa seperti *case folding*, *punctuation removal*, dan penghapusan *stopword.* Namun, daftar *stopword* yang digunakan berasal dari *library* NLTK. Pada teks berbahasa inggris, lemmatization juga digunakan untuk mengubah kata ke bentuk dasarnya. Misalkan, kata "*selection*" diubah menjadi "*select*" [21]. Dalam *preprocessing*, digunakan deketsi bahasa dengan *library* lengdetect untuk memproses sesuai dengan bahasanya. Langkah-langkah dalam preprocessing sangat penting untuk memastikan kualitas data lebih akurat yang akan di analisis lebih lanjut menggunakan LDA.

2.3 LDA Topic Modeling

Metode LDA *(Latent Dirichlet Allocation)* digunakan untuk menentukan distribusi topik dari judul dan abstrak penelitian. LDA adalah model generatif yang memiliki kemampuan untuk menemukan topik dalam dokumen tanpa terikat pada domain atau bahasa [12]. Dalam penelitian ini, LDA dibangun dengan parameter yang telah disesuaikan melalui uji coba menggunakan skor koherensi. Uji coba dengan berbagai iterasi untuk menemukan topik yang relevan dan memiliki keselarasan yang tinggi. Distribusi kata dalam dokumen dihitung menggunakan Bag of Words. Selanjutnya, distribusi ini membantu model LDA dalam menghasilkan topik tersembunyi, seperti kata kunci atau frasa yang sering muncul. Integrasi LDA dengan preferensi mahasiswa dilakukan melalui pencocokan topik yang dihasilkan dengan minat penelitian mahasiswa. Proses ini melibatkan analisis tambahan, seperti pembobotan nilai untuk menilai relevansi. Dengan pendekatan ini, LDA mendukung pengambilan keputusan dalam sistem rekomendasi dosen pembimbing secara efektif, memastikan topik penelitian yang dihasilkan dapat disesuaikan dengan kebutuhan mahasiswa [13][9].

2.4 TF-IDF

TF-IDF *(Term Frequency-Inverse Document Frequency)* merupakan metode statistik yang menggabungkan dua pendekatan untuk mengukur bobot suatu kata dalam dokumen.

$$tf = \frac{t}{n} \tag{1}$$

$$idf = log\frac{N}{df} \tag{2}$$

$$tf\ .\ idf = tf \times idf \tag{3}$$

tf adalah frekuensi kemunculan suatu kata (t) dalam dokumen, dan n adalah jumlah total kata dalam dokumen. Persamaan (1) digunakan untuk menghitung seberapa sering kata muncul dalam dokumen. Persamaan (2) mendefinisikan idf, yaitu rasio antara jumlah total dokumen (N) dengan jumlah dokumen yang mengandung kata tersebut (df) [22].

2.5 BERT

BERT *(Bidirectional Encoder Representations from Transformers)* merupakan teknik *self-attention* yang memanfaatkan bagian encoder dari arsitektur *transformer* [7][9]. BERT dapat memahami teks lebih baik dengan mempertimbangkan kata-kata di sebelah kiri dan kanan dari kata yang dianalisis. Untuk setiap kata dalam kalimat, model ini membuat representasi vektor yang menunjukkan hubungan dan konteksnya. Model rekomendasi menggunakan model BERT secara luas [23].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499

2.6 Cosine Similarity

Algoritma *Cosine Similarity* digunakan untuk menentukan seberapa mirip kedua dokumen [6]. Metode ini menghitung derajat kesamaan antara dua objek dalam vektor menggunakan kata kunci dari dokumen [19]. Jika dokumen identik, sudutnya adalah nol derajat dan kesamaannya satu. Sebaliknya, ketika kedua dokumen tidak sama sekali identik, sudutnya adalah 90° dan kesamaannya adalah nol [11].

$$Similarity = \cos(\theta)\ \frac{A.B}{\|A\|\|B\|} \tag{4}$$

2.7 Evaluasi

Untuk model rekomendasi dosen pembimbing dapat dievaluasi menggunakan perhitungan akurasi. Akurasi menunjukkan proporsi rekomendasi yang benar dibandingkan dengan total percobaan.

$$Akurasi = \frac{Jumlah\ percobaan\ relevan}{Total\ percobaan} \tag{5}$$

Penggunaan akurasi memudahkan perbandingan performa model rekomendasi. Nilai akurasi yang lebih tinggi menunjukkan rekomendasi yang lebih relevan. Selain itu juga penggunaan presisi, recall, dan F1-Scroe dilakukan. Presisi dilakukan untuk memastikan bahwa rekomendasi yang diberkan relevan, sementara recall untuk mengukur seberapa banyak item relevan yang berhasil terdeteksi sistem. F1-Score menggabungkan presisi dan recall menjadi satu nilai yang memberikan evaluasi yang seimbang antara keduanya [24].

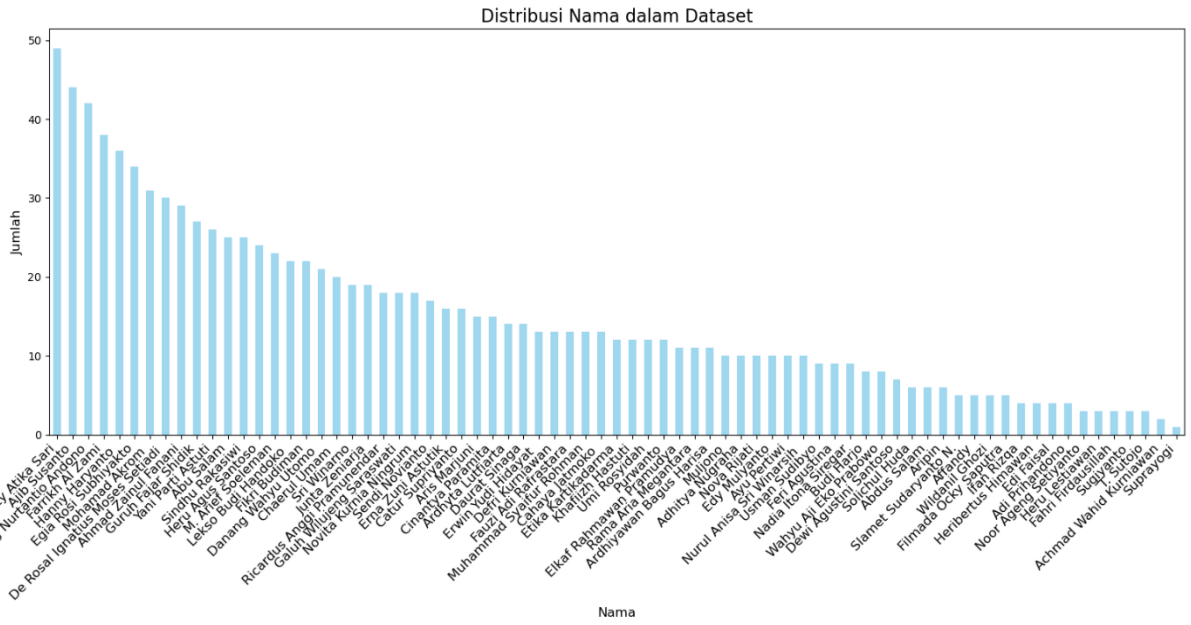
$$Presisi = \frac{Jumlah\ rekomendasi\ relevan}{5\ rekomendasi\ yang\ ditampilkan} \tag{6}$$

$$Recall = \frac{Jumlah\ rekomendasi\ relevan}{Total\ item\ relevan} \tag{7}$$

$$F1 - Score = 2 \cdot \frac{Presisi \cdot Recall}{Presisi + Recall} \tag{8}$$

3. HASIL DAN PEMBAHASAN

Data yang berhasil dikumpulkan dari 71 dosen dari Fakultas Teknik Informatika yaitu 1.096 baris data. Data ini dimulai dari tahun 2019-2024 dengan menggunakan teknik *scrapping*. Setelah proses pengumpulan selesai, hasil yang diperoleh disimpan dalam format CSV, untuk mempermudah proses analisis lebih lanjut. Gambar 2 menunjukkan hasil distribusi nama yang ada dalam dataset yang telah dikumpulkan, yang menggambarkan bagaimana data tersebar berdasarkan nama dosen.



Gambar 2. Hasil distribusi nama dalam dataset yang dikumpukan menggunakan teknik scrapping

Dalam penelitian ini, proses analisis teks dimulai dengan tahap preprocessing data untuk mengolah teks menjadi bersih dan terstruktur. Tahap ini sangat penting karena teks mentah yang diambil dari sumber data sering kali mengandung dokumen yang tidak relevan, seperti tanda baca, angka, atau kata-kata umum yang tidak memberikan informaasi untuk analisis lebih lanjut. Fokus utama dari tahap *preprocessing* ini hanya pada tiga kolom, yaitu nama,

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499

judul, dan abstrak. Ketiga kolom ini digunakan karena mengandung informasi yang paling relevan dengan analisis topik dan pemodelan yang dilakukan di tahap selanjutnya. Kolom judul dan abstrak digabungkan menjadi satu kolom untuk menyederhanakan proses analisi, sementara kolom nama digunakan sebagai label dalam analisis. Hasil dari *preprocessing* berupa teks yang telah terstruktur dan bersih dan siap untuk diproses lebih lanjut menggunakan teknik analisis topik. Tabel 1 menunjukan contoh dokumen sebelum dan sesudah tahap *preprocessing*.

Tabel 1. Hasil preprocessing dokumen

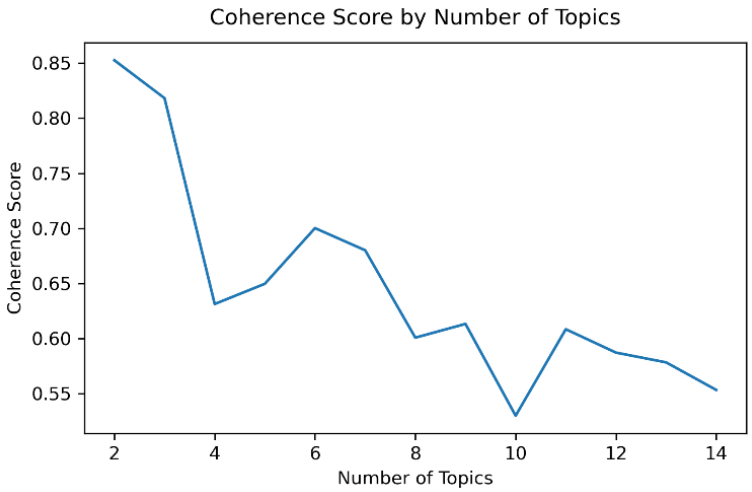
Sebelum *Preprocessing*		Sesudah *Preprocessing*
Judul	Abstrak	
Implementasi Algoritma K-Means Dalam Pengklasteran untuk Rekomendasi Penerima Beasiswa PPA di UDINUS	Rekomendasi penerima beasiswa Peningkatan Prestasi Akademik (PPA) dikelompokkan menjadi 2 cluster yaitu diterima dan tidak diterima untuk mendapatkan beasiswa. Algoritma K-Means merupakan teknik unsupervised learning yang dapat digunakan dalam mengelompokkan data pengajuan beasiswa. Tujuan dari penelitian ini adalah untuk merekomendasikan penerima beasiswa dengan menggunakan algoritma k-means, hasil rekomendasi berupa penempatan data pendaftar beasiswa ke masing-masing kelompok cluster yang dihasilkan. Eksperimen proses clustering dilakukan menggunakan data pendaftar beasiswa PPA dari biro kemahasiswaan udinus tahun 2016 sebanyak 44 pendaftar beasiswa PPA. Melalui seleksi atribut, k-means ini melakukan perhitungan untuk menempatkan setiap data ke cluster yang sudah ditentukan. Sebanyak 154 mahasiswa direkomendasikan mendapatkan beasiswa PPA sedangkan 287 mahasiswa tidak mendapatkan.	Implementasi algoritma kmeans pengklasteran rekomendasi terima beasiswa ppa udinus rekomendasi terima beasiswa tingkat prestasi akademik ppa kelompok cluster terima terima beasiswa algoritma kmeans teknik unsupervised learning kelompok data aju beasiswa tuju teliti rekomendasi terima beasiswa algoritma kmeans hasil rekomendasi tempat data daftar beasiswa masingmasing kelompok cluster hasil eksperimen proses clustering data daftar beasiswa ppa biro mahasiswa udinus daftar beasiswa ppa seleksi atribut kmeans hitung tempat data cluster tentu mahasiswa rekomendasi beasiswa ppa mahasiswa
Evaluation of feature selection using wrapper for numeric dataset with random forest algorithm	Preprocessing is more than half of machine learning process. Dimensionality reduction is one of the preprocessing task, which included feature extraction and selection. Feature selection used for identify relevant and remove not relevant feature. The goal of this research is to select relevant feature using wrapper method for early diabetes prediction dataset which has been transformed to numeric dataset previously. Forward and backward selection are used in wrapper method, that's combine with random forest and cross validation. Random forest is decision tree enhancement, which is group of trees that can produce difference or same result at each tree. The most results are made as final result. The final result from feature selection with wrapper method can make higher accuracy than without feature selection for numeric dataset and the number of feature can be reduced. With features selection which is sequential forward selection it has 98.84 % accuracy with 11 feature selected and with sequential backward selection, it has 99.03 % accuracy with same number of features	evaluation feature selection using wrapper numeric dataset random forest algorithm preprocessing half machine learning process dimensionality reduction one preprocessing task included feature extraction selection feature selection used identify relevant remove relevant feature goal research select relevant feature using wrapper method early diabetes prediction dataset transformed numeric dataset previously forward backward selection used wrapper method thats combine random forest cross validation random forest decision tree enhancement group trees produce difference result tree results made final result final result feature selection wrapper method make higher accuracy without feature selection numeric dataset number feature reduced features selection sequential forward selection accuracy feature selected sequential backward selection accuracy number features selected reduced features reduces complexity trees time required mining process

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499

Sebelum *Preprocessing*		Sesudah *Preprocessing*
Judul	Abstrak	
	selected. With reduced features, will reduces complexity of trees and time required in mining process.	

Tabel 1 menujukkan perbedaan yang terjadi pada dokumen sebelum dan sesudah tahap *preprocessing*. Dalam kolom sebelum preprocessing, teks mengandung banyak kata yang tidak relevan seperti “dalam”, “untuk”, dan lainnya. Setelah dilakukan *preprocessing*, teks menjadi lebih berfokus dan berisi kata-kata kunci. Proses ini melibatkan pengubahan huruf kapital menjadi huruf kecil, penghapusan kata-kata umum *(Stopwords)*, *stemming* untuk mengubah kata ke bentuk dasar, serta penghapusan tanda baca dan elemen yang tidak relevan, serta penggabungan kolom judul dan abstrak. Hasil preprocessing tidak hanya membuat teks lebih bersih tetapi juga meningkatkan kualitas data untuk analisis lebih lanjut.

Setelah tahap preprocessing teks, langkah berikutnya adalah menerapkan pemodelan topik menggunakan metode *Latent Dirichlet Allocation* (LDA). Metode ini bertujuan untuk mengidentifikasi tema-tema yang tersembunyi dalam kumpulan dokumen, yang disebut “topik”. Dalam penelitian ini, diputuskan untuk mengatur jumlah topik yang dihasilkan oleh *Latent Dirichlet Allocation* menjadi dua berdasarkan perhitungan skor koherensi, yang digunakan untuk mengevaluasi sejauh mana kumpulan dokumen dapat dikelompokkan ke dalam topik yang relevan. Skor koherensi memberikan indikasi bahwa pengelompokan dokumen ke dalam dua topik mampu menggambarkan konten dengan lebih baik dan relevan. Pemilihan jumlah topik merupakan langkah penting, karena terlalu sedikit atau terlalu banyak topik dapat mengakibatkan kehilangan informasi dan pengelompokan dokumen yang kurang optimal.



Gambar 3. Hasil skor koherensi LDA

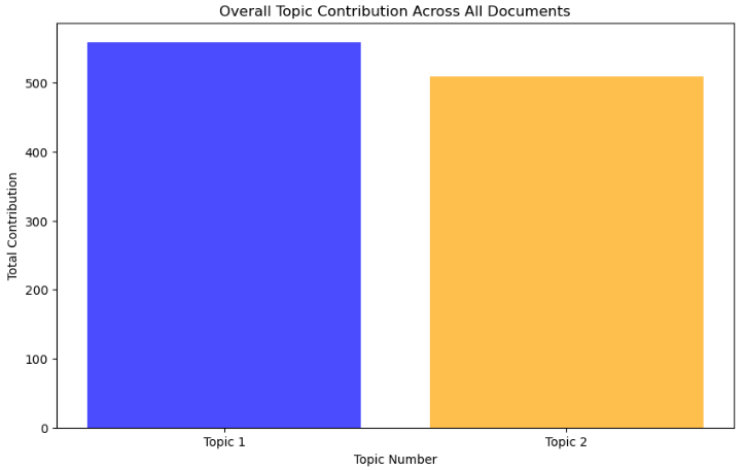
Dalam Gambar 2, terlihat bahwa pemilihan jumlah dua topik menghasilkan skor koherensi tertinggi, yang menunjukkan bahwa pembagian ke dalam dua kategori menghasilkan kelompok yang lebih informatif dan relevan. Setelah pemilihan jumlah topik, langkah selanjutnya adalah menganalisis distribusi istilah dalam masing-masing topik yang dihasilkan oleh LDA. Tabel 2 menunjukkan kata kunci yang mewakili masing-masing topik, dimana setiap topik direpresentasikan melalui kombinasi kata-kata dengan bobot tertentu. Bobot ini menggambarkan tingkat relevasi kata-kata tersebut terhadap topik yang bersangkutan.

Tabel 2. Hasil topik LDA

Topik	Topik Tersembunyi
1	0.011*”data” + 0.008*”using” + 0.006*”model” + 0.005*”method” + 0.005*”ajar” + 0.005*”tingkat” + 0.005*”study” + 0.005*”hasil” + 0.004*”used” + 0.004*”metode”
2	0.008*”data” + 0.006*”using” + 0.006*”method” + 0.006*”image” + 0.005*”hasil” + 0.005*”research” + 0.005*”teliti” + 0.005*”model” + 0.005*”learning” + 0.004*”used”

Berdasarkan distribusi kata kunci, bahwa topik pertama berfokus pada istilah-istilah seperti “data”, “method”, dan “model”, yang cenderung berkaitan dengan proses dan metodologi penelitian. Sementara itu, topik kedua lebih menekankan “image”, “research”, dan “learning”, yang menunjukkan pada studi berbasis gambar dan pembelajaran.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499



Gambar 4. Pembagian topik LDA pada dataset

Gambar 3 menggambarkan pembagian dataset terhadap kedua topik oleh LDA. Kontribusi kedua topik yang dihasilkan oleh *Latent Dirichlet Allocation* cukup seimbang, yang menandakan bahwa model *Latent Dirichlet Allocation* mampu membagi dokumen menjadi dua topik dengan proporsi yang hampir merata.

Setelah pemodelan topik menggunakan *Latent Dirichlet Allocation*, selanjutnya penerapan teknik TF-IDF *(Term Frequency-Inverse Document Frequency)* dan BERT *(Bidirectional Encder Representasions form Transformers)* untuk menghasilkan embedding representasi dari dokumen. Proses ini memberikan vektor yang kaya akan informasi untuk setiap dokumen yang digunakan dalam pengukuran kesamaan. Setelah mendapatkan hasil vektor. Setelah mendapatkan vektor representasi setiap dokumen, dilakukan pencarian tingkat kemiripan dokumen dengan menggunakan *Cosine Similarity*. *Cosine similarity* mengukur kemiripan vektor dengan data testing dan data training pada penelitian.

Sebagai contoh, untuk lima dokumen model ini memberikan lima rekomendasi teratas yang relevan berdasarkan algoritma *Cosine Similarity*. Hasil rekomendasi tersebut disajikan pada Tabel 3.

Tabel 3. Hasil rekomendasi Cosine Similarity

Dokumen	Judul	Nama dosen	*Similarity*
1	“EVALUASI PERFORMA TEKNIK SAMPLING DALAM MENANGANI IMBALANCE DATA PADA KLASIFIKASI DIABETES”	[AB]	0.957
		[MSR]	0.956291
		[JZ]	0.95495
		[AL]	0.954355
		[AN]	0.954212
2	“IMPLEMENTASI ALGORITMA RANDOM FOREST DALAM KLASIFIKASI KETEPATAN WAKTU KELULUSAN MAHASISWA TEKNIK INFORMATIKA UNIVERSITAS DIAN NUSWANTORO”	[BH]	0.959
		[AN]	0.956555
		[GWS]	0.956314
		[CP]	0.956279
		[FAZ]	0.954563
3	“ANALISIS SENTIMEN MASYARAKAT TENTANG VIRTUAL YOUTUBER PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE ALGORITMA NAIVE BAYES”	[SW]	0.963
		[NASW]	0.959763
		[RAP]	0.956958
		[AS]	0.95631
		[AL]	0.955981
4	“Prediksi Hasil Panen Ikan Lele untuk Pemenuhan Kebutuhan Masyarakat Menggunakan Metode Algoritma Multiple Linier Regression”	[SN]	0.944
		[SW]	0.938435
		[YPA]	0.937872
		[ERS]	0.937835
		[HH]	0.937373
5	“Deteksi Jamur Roti Tawar Menggunakan K-Means Clustering Berdasarkan Grey Level Co-Occurrence Matrix dan Region of Interest”	[RAM]	0.961
		[RAP]	0.960867
		[AS]	0.959562
		[AN]	0.956778
		[CAS]	0.956058

Hasil pengujian menunjukan bahwa nilai similarity diatas 0.95 dianggap berhasil dalam mengidentifikasi hubungan antara preferensi pengguna dengan karakteristik item yang tersedia di dataset. Dari lima percobaan dengan menggunakan dokumen yang berbeda, diperoleh rata-rata nilai similarity sebesar 0.9536. nilai rata-rata ini menunjukkan efektivitas metode *Cosine Similarity* dalam menemukan kesamaan dokumen yang relevan.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499

Untuk mengevaluasi performa model secara keseluruhan, pengujian dilakukan dengan memberikan lima rekomendasi nama dosen dengan pencocokan menggunakan *Cosine Similarity* seperti pada Tabel 3, yang kemudian dibandingkan dengan daftar dosen relevan yang seharusnya direkomendasikan.

Tabel 4. Hasil rekomendasi dosen pembimbing

Uji coba	Top-5 rekomendasi dosen	Dosen relevan	Jumlah	Keterangan
1	[AB], [MSR], [JZ], [AL], [AN]	[AB], [JZ], [AL], [AN]	4	Valid
2	[BH], [AN], [GWS], [CP], [FAZ]	[AN], [FAZ]	2	Tidak Valid
3	[SW], [NASW], [RAP], [AS], [AL]	[SW], [NASW], [RAP], [AL]	4	Valid
4	[AN], [SW], [YPA], [ERS], [HH]	[SN], [SW], [YPA]	3	Valid
5	[RAM], [RAP], [AS], [AN], [CAS]	[RAP], [CAS]	2	Tidak Valid
6	[AS], [EZA], [ERP], [AN], [UR]	[AS], [ERP], [AN]	3	Valid
7	[RAP], [ERS], [EZA], [FA], [NKN]	[ERS], [FA], [NKN]	3	Valid
8	[AN], [AL], [CP], [DWU], [NASW]	[AN], [AL], [CP], [DWU]	4	Valid
9	[FB], [HH], [AS], [S], [NKN]	[FB], [AS], [S], [NKN]	4	Valid
10	[ERP], [GWS], [MSR], [AS], [JZ]	[GWS], [MSR], [AS]	3	Valid

Berdasarkan Tabel 4, rekomendasi dianggap valid jika jumlah dosen relevan tiga atau lebih, sedangkan jika jumlah dosen relevan kurang dari tiga, maka dianggap tidak valid. Secara keseluruhan, hasil pengujian menunjukkan bahwa model mampu memberikan rekomendasi yang cukup baik, dengan akurasi sebesar 80%. Hal ini menunjukkan bahwa model yang digunakan, yaitu kombinasi LDA, TF-IDF, dan BERT memiliki kemampuan yang cukup efektif dalam dalam memberikan rekomendasi dosen pembimbing yang relevan dan tepat.

Meskipun akurasi menunjukkan bahwa model cukup efektif dalam memberikan rekomendasi yang relevan dan tepat, hasil evaluasi presisi dan recall menunjukkan adanya perbaikan. Presisi sebesar 66% menunjukkan bahwa sebagian besar rekomendasi yang diberikan relevan, namun recall hanya 19% menujukkan bahwa banyak item relevan yang terlewatkan. Hal ini dikarenakan sistem hanya menampilkan lima rekomendasi teratas saja, meskipun sebenarnya ada lebih banyak item relevan yang disarankan. F1-score menghasilkan nilai sekitar 29%, meskipun hasil presisi cukup baik, rendahnya recall membuat keseimbangan antara keduanya perlu diperbaiki.

Keberhasilan ini menunjukkan potensi dari pendekatan yang diterapkan dalam penelitian ini untuk mendukung proses pemilihan dosen pembimbing yang lebih efisien dan akurat. dengan metode ini, mahasiswa dapat menemukan dosen yang sesuai dengan minat penelitian mereka, sementara dosen juga dapat menerima mahasiswa yang memiliki kesesuaian dengan bidang keahliannya. Penelitian ini menunjukkan bahwa penerapan teknologi dalam dunia pendidikan dapat memberikan manfaat yang signifikan, khususnya dalam proses akademik seperti pemilihan dosen pembimbing.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa rekomendasi dosen pembimbing berbasis *text similarity* dengan pemodelan menggunakan LDA, TF-IDF, dan BERT berhasil memberikan rekomendasi yang relevan dan akurat. Proses *preprocessing* data, seperti penggabungan kolom judul dan abstrak, serta pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA), terbukti penting dalam menghasilkan dataset yang terstruktur dan representasi yang informatif. Skor koherensi yang diperoleh dalam pemodelan LDA menujukkan bahwa pembagian topik menjadi dua kategori menghasilkan pemahaman yang lebih jelas terhadap dokumen yang dianalisis. Selanjutnya, teknik TF-IDF dan BERT digunakan untuk menghasilkan vektor representasi dokumen yang kemudian diproses menggunakan metode *Cosine Similarity* untuk mencocokan dokumen yang relevan. Pengujian dengan lima dokumen menunjukkan nilai rata-rata similarity sebesar 0.9536, yang menandakan efektivitas metode *Cosine Similarity* dalam menemukan dokumen yang relevan. Hasil ini menunjukkan kemampuan yang baik dalam mencocokkan topik penelitian mahasiswa dengan bidang keahlian dosen. Uji coba lebih lanjut menunjukkan nilai akurasi sebesar 80%. Meskipun demikian, analisis lebih lanjut menunjukkan adanya tantangan dalam meningkatkan *recall* yang rendah, dengan banyak item relevan yang tidak terdeteksi. Hal ini mengindikasikan bahwa meskipun rekomendasi yang diberikan relevan dilihat dari nilai presisi, masih ada banyak potensi rekomendasi yang terlewatkan. Keberhasilan dalam memberikan rekomendasi menunjukkan potensi penerapan teknologi berbasis text analysis dalam mendukung proses pemilihan dosen pembimbing yang lebih efisien dan tepat sasaran. Dengan pemodelan ini, mahasiswa dapat dengan mudah memilih dosen pembibing yang sesuai dengan topik penelitian mereka, sedangkan dosen dapat menerima mahasiswa yang memiliki kesesuaian dengan bidang keahliannya. Dengan demikian, model yang diusulkan diharapkan dapat memperbaiki proses akademik di perguruan tinggi melalui penggunaan teknologi berbasis analisis teks.

REFERENCES

[1] R. Megawati and M. Damayanti, “Peran Dosen Pembimbing Skripsi dalam Proses Penyelesaian Tugas Akhir Mahasiswa,” *Journal of Health, Education, Economics, Science, and Technology (J-HEST)* , vol. 4, pp. 33–39, 2021

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2175−2183
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6499

[2] R. Ario Nugroho, "Analisa Penentuan Dosen Pembimbing Tugas Akhir Mahasiswa Menggunakan Naive Bayes Classifier," *Jurnal SIMTIKA*, vol. 4, no. 3, 2021.

[3] N. Andriani and B. Wibawanta, "Peran Dosen Pembimbing Sebagai Pemimpin Yang Melayani Dalam Pembimbingan Tugas Akhir Mahasiswa Program Sarjana [The Role Of Supervisor As A Servant Leader In The Final Project Supervision Of Undergraduate Students]," *Polyglot: Jurnal Ilmiah*, vol. 16, no. 2, pp. 230–251, Jun. 2020, doi: 10.19166/pji.v16i2.1927.

[4] D. Aqmala, I. Farida, A. S. Samasta, and A. Setiawan, "Kompetensi Kinerja Dosen Terhadap Topik Bimbingan Tugas Akhir Mahasiswa Menggunakan Naïve Bayes," *TRANSFORMTIKA*, vol. 19, no. 1, pp. 48–56, 2021.

[5] H. Hairani and M. Mujahid, "Recommendations of Thesis Supervisor using the Cosine Similarity Method," *SISTEMASI*, vol. 11, no. 3, p. 646, Sep. 2022, doi: 10.32520/stmsi.v11i3.2003.

[6] A. Azhari, E. Buulolo, and N. Silalahi, "Sistem Rekomendasi Dosen Pendamping Skripsi Berbasis Text Rank menggunakan Metode Cosine Similarity," *Pelita Informatika : Informasi dan Informatika*, vol. 10, no. 3, pp. 119–122, 2022.

[7] S. Wehnert, V. Sudhi, S. Dureja, L. Kutty, S. Shahania, and E. W. De Luca, "Legal norm retrieval with variations of the bert model combined with TF-IDF vectorization," in *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021*, Association for Computing Machinery, Inc, Jun. 2021, pp. 285–294. doi: 10.1145/3462757.3466104.

[8] J. W. Sun, J. Q. Bao, and L. P. Bu, "Text Classification Algorithm Based on TF-IDF and BERT," in *Proceedings - 2022 11th International Conference of Information and Communication Technology, ICTech 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 533–536. doi: 10.1109/ICTech55460.2022.00112.

[9] N. Yang, J. Jo, M. Jeon, W. Kim, and J. Kang, "Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models," in *Expert Systems with Applications*, Elsevier Ltd, Mar. 2022. doi: 10.1016/j.eswa.2021.116209.

[10] Z. Jin, X. Lai, and J. Cao, "Multi-label Sentiment Analysis Base on BERT with modified TF-IDF," in *ISPCE-CN 2020 - IEEE International Symposium on Product Compliance Engineering-Asia 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ISPCE-CN51288.2020.9321861.

[11] E. Rivadeneira-Pérez and C. Callejas-Hernández, "Leveraging LDA Topic Modeling and BERT Embeddings for Thematic Unsupervised Classification of Tourism News in Rest-Mex Competition," *IberLEF*, vol. 1, 2023, [Online]. Available: http://ceur-ws.org

[12] Y. Zhang and L. Zhang, "Movie Recommendation Algorithm Based on Sentiment Analysis and LDA," in *Procedia Computer Scienc2*, Elsevier B.V., 2022, pp. 871–878. doi: 10.1016/j.procs.2022.01.109.

[13] P. N. Andono, Sunardi, R. A. Nugroho, and B. Harjo, "Aspect-Based Sentiment Analysis for Hotel Review Using LDA, Semantic Similarity, and BERT," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 5, pp. 232–243, Oct. 2022, doi: 10.22266/ijies2022.1031.21.

[14] D. Jurafsky and J. H. Martin, *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2020.

[15] Denis Rothman, *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*, 2nd ed. Birmingham, Inggris, 2022.

[16] A. Rahmatulloh and R. Gunawan, "Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar," *Indonesian Journal of Information Systems (IJIS*, vol. 2, no. 2, 2020, [Online]. Available: http://garuda.ristekdikti.go.id/,

[17] N. Adila, "Implementation of Web Scraping for Journal Data Collection on the SINTA Website," *Sinkron*, vol. 7, no. 4, pp. 2478–2485, Oct. 2022, doi: 10.33395/sinkron.v7i4.11576.

[18] Y. Adilaksa and A. Musdholifah, "Recommendation System for Elective Courses using Content-based Filtering and Weighted Cosine Similarity," in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 51–55. doi: 10.1109/ISRITI54043.2021.9702788.

[19] R. Al Rasyid, D. Handayani, and U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, pp. 171–177, 2024, doi: 10.35870/jti.

[20] E. M. Sipayung, "Sentiment on Public Trust Using the NLP Rule Based Method," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 12, no. 1, pp. 175–182, Jan. 2024, doi: 10.26418/justin.v12i1.72426.

[21] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ Res Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.

[22] R. Rismanto, A. R. Syulistyo, and B. P. C. Agusta, "Research supervisor recommendation system based on topic conformity," *International Journal of Modern Education and Computer Science*, vol. 12, no. 1, pp. 26–34, 2020, doi: 10.5815/ijmecs.2020.01.04.

[23] C. Jeong, S. Jang, H. Shin, E. Park, and S. Choi, "A Context-Aware Citation Recommendation Model with BERT and Graph Convolutional Networks," in *The Association Of Computational Linguistics Anthology Network*, Mar. 2019. [Online]. Available: http://arxiv.org/abs/1903.06464

[24] Z. Fayyaz, M. Ebrahimian, D. Nawara, A. Ibrahim, and R. Kashef, "Recommendation systems: Algorithms, challenges, metrics, and business opportunities," *Applied Sciences (Switzerland)*, vol. 10, no. 21, pp. 1–20, Nov. 2020, doi: 10.3390/app10217748.