Jurnal Ilmu Komputer Agri-Informatika

Tersedia secara daring di:
https://jurnal.ipb.ac.id/index.php/jika

Volume 13 Nomor 1 halaman 44–55
eISSN: 2654-9735, pISSN: 2089-6026

Penerapan *Information Retrieval System* dan *Latent Dirichlet Allocation* sebagai *Service* untuk Pemetaan SDGs

Implementation of Information Retrieval System and Latent Dirichlet Allocation as a Service for SDGs Mapping

DARREN ALEXANDER[1], KARLISA PRIANDANA[1]*, SHELVIE NIDYA NEYMAN[1], JULIO ADISANTOSO[1]

Abstrak

Penentuan kategori *Sustainable Development Goals* (SDGs) pada data penelitian dan pengabdian kepada masyarakat (PPM) di IPB *University* pada periode awal dilakukan secara manual, sehingga sebagian besar data belum memiliki label SDGs dan menimbulkan ketidakkonsistenan informasi. Penelitian ini mengusulkan pendekatan otomatis untuk pelabelan SDGs berbasis *Information Retrieval System* (IR *System*) yang terintegrasi dengan KMS PPM IPB. Kontribusi utama penelitian ini adalah pengembangan *framework* otomasi yang menggabungkan pemodelan topik *Latent Dirichlet Allocation* (LDA) dan *semantic similarity* berbasis *triplet loss* untuk mengukur kesesuaian semantik antara dokumen PPM dan kategori SDGs. Selain itu, sistem dirancang sebagai layanan berbasis REST API yang memungkinkan integrasi langsung dengan sistem KMS PPM IPB, sehingga mendukung implementasi dalam lingkungan sistem nyata dan meningkatkan skalabilitas serta interoperabilitas. Hasil evaluasi menunjukkan nilai *F1-score* sebesar 16%, yang mengindikasikan keterbatasan dalam menyeimbangkan *precision and recall*. Namun demikian, rata-rata waktu *respons* sebesar 5,48 detik menunjukkan bahwa sistem mampu beroperasi dalam batas performa yang dapat diterima. Hasil ini menunjukkan bahwa pendekatan yang diusulkan dapat menjadi tahap awal dalam otomatisasi pelabelan SDGs untuk pengelolaan data PPM secara terintegrasi.

Kata Kunci**:** *information retrieval system, latent dirichlet allocation, REST API, SDGs, semantic similarity, triplet loss*

Abstract

The assignment of Sustainable Development Goals (SDGs) categories to research and community service (PPM) data at IPB University was initially conducted manually, resulting in a large portion of data lacking SDG labels and causing inconsistencies. This study proposes an automated SDG labeling approach using an Information Retrieval System (IR System) integrated with the KMS PPM IPB. The main contribution lies in the development of an automation framework that combines Latent Dirichlet Allocation (LDA) for topic modeling and semantic similarity based on triplet loss to measure semantic relevance between PPM documents and SDG categories. Furthermore, the system is implemented as a REST API-based service, enabling seamless integration with the KMS environment, and supporting deployment in a real-world system context with improved scalability and interoperability. Evaluation results show an F1-score of 16%, indicating limitations in balancing precision and recall. However, the average response time of 5.48 seconds demonstrates acceptable performance for an integrated system. These findings suggest that the proposed approach can serve as an initial step toward automated SDG labeling in large-scale PPM data management.

Keywords: information retrieval system, knowledge management system, latent dirichlet allocation (LDA), semantic similarity, sustainable development goals (SDGs), triplet loss

[1]Program Studi Ilmu Komputer, Sekolah Sains Data, Matematika, dan Informatika, Institut Pertanian Bogor, Bogor 16680;
*Penulis Korespondensi: Tel/Faks: 0251-8625584; Surel: karlisa@apps.ipb.ac.id;

PENDAHULUAN

Perserikatan Bangsa-Bangsa mengadopsi *Sustainable Development Goals* (SDGs) pada 25 September 2015 untuk mengatasi berbagai tantangan global dan mendorong pembangunan berkelanjutan di seluruh dunia. Sejak diadopsi, minat dari kalangan bisnis, organisasi, dan pemerintah dalam menggunakan SDGs sebagai kerangka pembangunan berkelanjutan terus meningkat (Rinaldi *et al.* 2024). Indonesia turut berperan aktif dalam implementasi SDGs melalui berbagai inisiatif pembangunan, termasuk di sektor pendidikan (Rulandari 2021). Komitmen ini tercermin dalam kebijakan strategis nasional, salah satunya Rencana Pembangunan Jangka Menengah Nasional (RPJMN) 2020–2024 yang menekankan peran perguruan tinggi sebagai agen perubahan dalam mendukung pencapaian SDGs.

IPB *University* merupakan salah satu perguruan tinggi yang berkontribusi dalam implementasi SDGs melalui kegiatan penelitian dan pengabdian kepada masyarakat (PPM). Sebagai bentuk dukungan, IPB *University* mengembangkan *Knowledge Management System* Penelitian dan Pengabdian kepada Masyarakat IPB (KMS PPM IPB) yang menyimpan informasi seluruh kegiatan PPM yang dilakukan oleh dosen, peneliti, pakar, dan mahasiswa. Sistem ini dimanfaatkan sebagai referensi akademik dan telah menghimpun sebanyak 9.344 data kegiatan PPM, termasuk informasi terkait SDGs. Namun, hasil analisis menunjukkan bahwa lebih dari 50% data PPM di KMS PPM IPB belum dilengkapi dengan informasi pemetaan SDGs. Oleh karena itu, diperlukan metode untuk memetakan kategori SDGs pada data yang belum memiliki informasi tersebut.

Pemetaan kategori SDGs telah banyak dilakukan dengan berbagai pendekatan. Hsu *et al.* (2022) mengembangkan teknik *Combinatorial Fusion Algorithm* (CFA) dengan mengombinasikan *Latent Dirichlet Allocation* (LDA) dan pendekatan *semantic link* untuk meningkatkan akurasi klasifikasi SDGs. Hajikhani dan Souminen (2022) memetakan SDGs pada dokumen sains dan paten menggunakan algoritma *machine learning* dengan *teknik natural language processing* (NLP). Selain itu, Guisiano *et al.* (2022) mengembangkan SDG Meter menggunakan *Bidirectional Encoder Representations from Transformers* (BERT). Rinaldi *et al.* (2024) menggunakan *classification algorithm* untuk mengklasifikasikan teks dalam dokumen organisasi ke dalam kategori SDGs.

Meskipun berbagai pendekatan ini telah dikembangkan, sebagian besar penelitian masih berfokus pada evaluasi model secara terpisah dan belum banyak diimplementasikan dalam sistem nyata yang terintegrasi. Selain itu, integrasi antar metode dalam satu *pipeline* untuk mendukung pemetaan SDGs secara otomatis masih relatif terbatas. Oleh karena itu, penelitian ini mengusulkan pengembangan sistem berbasis layanan (*service*) untuk pemetaan SDGs berdasarkan informasi dokumen.

Penelitian ini mengimplementasikan *Information Retrieval System* (IR *System*) untuk mengidentifikasi informasi SDGs yang relevan dengan memanfaatkan abstrak laporan kegiatan PPM. IR *System* telah banyak digunakan dalam pencarian informasi. Adiyanto dan Handayani (2022) menerapkan IR *System* untuk pencarian dokumen menggunakan metode *Vector Space Model* (VSM), sedangkan Makmum *et al.* (2022) menggunakan metode yang sama untuk pencarian artikel arkeologi. Pal dan Mukhopadhyay (2024) menerapkan beberapa *machine learning* seperti model klasik *Support Vector Machine* (SVM), *Random Forest*, dan *Naive Bayes* untuk memetakan SDGs pada *Indian Research Publications.*

Pendekatan *Information Retrieval* tradisional seperti VSM masih bergantung pada kesamaan leksikal antara *query* dan dokumen, sehingga memiliki keterbatasan dalam menangkap hubungan semantik. Berbagai metode berbasis NLP telah dikembangkan untuk mengatasi keterbatasan ini melalui representasi semantik, seperti penggunaan *embedding* dan pengukuran *semantic similarity*. Salah satu pendekatan yang banyak digunakan adalah *sentence embedding* menggunakan kerangka *Sentence Transformers* (Reimers dan Gurevych 2019). Namun, pendekatan tersebut umumnya menerapkan *semantic similarity* secara langsung pada representasi dokumen atau kalimat tanpa mempertimbangkan struktur topik laten. Selain itu, penerapan teknik ini membutuhkan sumber daya komputasi yang besar (Lin *et al.* 2021).

Beberapa pendekatan lainya pada pemetaan SDGs masih menggunakan representasi teks berbasis TF-IDF atau *bag-of-words* yang dikombinasikan dengan model klasifikasi tradisional seperti *Support Vector Machine* (SVM), yang cenderung belum mampu menangkap konteks semantik dan hubungan antar kata secara mendalam.

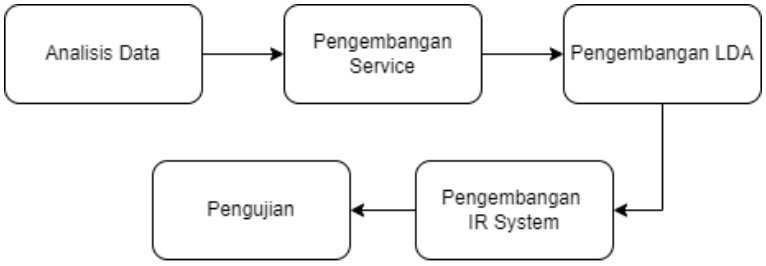
Oleh karena itu, penelitian ini mengusulkan integrasi pemodelan topik dan *semantic similarity* dalam satu *pipeline*, di mana distribusi topik hasil *Latent Dirichlet Allocation* (LDA) digunakan sebagai representasi sebelum dilakukan pengukuran *semantic simmilarity* berbasis *triplet loss*. Pendekatan ini bertujuan agar proses pengukuran kesamaan tidak hanya mempertimbangkan makna teks secara langsung, tetapi juga didasarkan pada struktur topik yang telah terbentuk, sehingga berpotensi meningkatkan kualitas pemetaan informasi.

LDA telah banyak digunakan untuk mengidentifikasi distribusi topik dalam dokumen. Karmila dan Ardianti (2022) menerapkan LDA untuk menentukan distribusi topik pada teks berita, sedangkan Sharma *et al.* (2022) mengembangkan *information modelling* menggunakan LDA untuk menganalisis tren dan pola penelitian rantai pasok dari berbagai basis data digital seperti IEEE, Springer, dan ACM. Selain itu, Li *et al.* (2025) menerapkan LDA untuk mengidentifikasi pusat penelitian dan evolusi pada analisis bibliometrik terhadap 1.366 artikel tentang PV terdistribusi yang diterbitkan di *Web of Science* selama tahun 1985–2023.

Kontribusi utama penelitian ini terletak pada integrasi distribusi topik sebagai representasi antara dalam pengukuran *semantic similarity* serta implementasinya dalam sistem berbasis layanan yang terintegrasi dengan KMS PPM IPB, membuat *IR System* dapat digunakan didalam implementasi sistem nyata.

METODE

Tahapan penelitian ini ditunjukkan pada Gambar 1, yang terdiri atas beberapa proses utama, yaitu analisis data, pengembangan *service*, pengembangan model LDA, pengembangan IR *System*, serta pengujian sistem.



Gambar 1 Tahapan penelitian

Analisis Data

Tahap awal penelitian dimulai dengan analisis data untuk memahami karakteristik data PPM yang akan digunakan, termasuk identifikasi kelengkapan informasi serta kebutuhan sistem. Data PPM diekstraksi langsung dari basis data pada tanggal 23 Oktober 2023 sebagai sumber data utama dalam penelitian ini. Berdasarkan hasil analisis yang telah dilakukan, terdapat sebanyak 9.344 data PPM yang tersimpan dalam basis data KMS IPB. Dari keseluruhan data tersebut, lebih dari 50% data belum dilengkapi dengan informasi SDGs yang relevan. Temuan ini menunjukkan adanya kesenjangan informasi yang signifikan, yang berpotensi menghambat proses pengelolaan dan pemanfaatan data PPM dalam konteks pemetaan kontribusi terhadap SDGs.

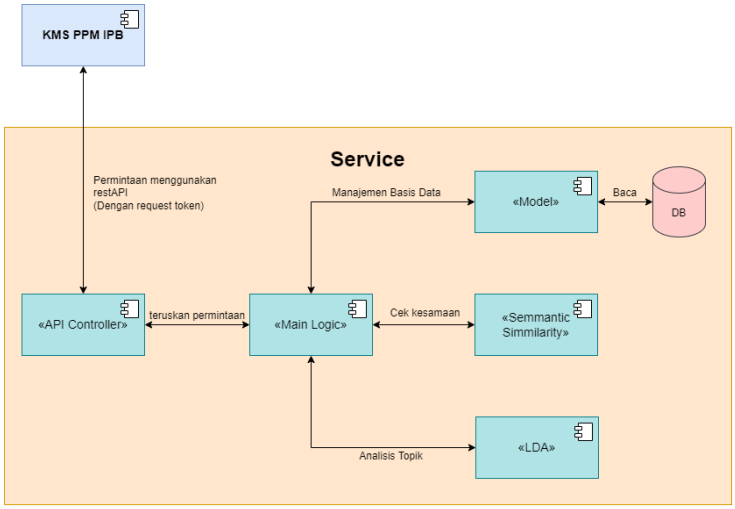
Selain itu, hasil analisis juga menunjukkan bahwa basis data KMS PPM IPB memiliki kumpulan *master data* yang menyimpan informasi lengkap terkait SDGs. *Master data* SDGs yang digunakan dalam penelitian ini berupa kumpulan informasi terkait 17 kategori *Sustainable Development Goals* yang diperoleh dari dokumen resmi Perserikatan Bangsa-Bangsa (*United Nations*). Data tersebut terdiri atas nama tujuan SDGs dan kata kunci yang merepresentasikan masing-masing kategori. Seluruh data kemudian disimpan dalam basis data dan kata kunci dari

SDGs digunakan sebagai sumber referensi dalam proses pemetaan *semantic similarity* terhadap dokumen PPM.

Pengembangan *Service*

Sistem dikembangkan menggunakan pendekatan berbasis layanan (*service-based architecture*), di mana setiap fungsi utama diimplementasikan sebagai komponen layanan yang terpisah dalam satu kesatuan sistem. Pendekatan ini memungkinkan pemisahan tanggung jawab antar komponen sehingga proses pengelolaan dan pengembangan sistem menjadi lebih terstruktur dan fleksibel. Studi terbaru menunjukkan bahwa arsitektur berbasis layanan mampu meningkatkan skalabilitas dan ketahanan sistem, serta mendukung integrasi yang lebih efisien antar komponen dalam lingkungan sistem yang kompleks (Li *et al.* 2021).

Seperti yang diilustrasikan pada Gambar 2, layanan (*service*) yang dikembangkan terdiri atas beberapa komponen utama, yaitu *API Controller* dan *Main Logic*. *API Controller* berfungsi sebagai antarmuka untuk menerima permintaan pencarian informasi melalui *HTTP request*. Selanjutnya, *Main Logic* bertindak sebagai komponen inti yang mengelola alur proses sistem, mulai dari pengambilan data SDGs dari basis data, pelaksanaan analisis topik melalui pemodelan LDA, hingga proses evaluasi kesamaan topik menggunakan algoritma *semantic similarity*.



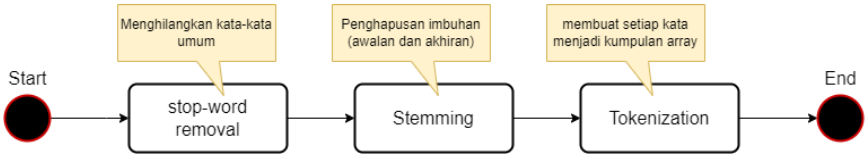
Gambar 2 Relasi Komponen *Service*

Dalam implementasinya, apabila terdapat data PPM yang belum memiliki informasi SDGs, sistem KMS PPM IPB akan mengirimkan parameter berupa dokumen abstrak ke *service* ini untuk diproses lebih lanjut. *Service* kemudian akan melakukan analisis terhadap data tersebut untuk mengidentifikasi informasi SDGs yang relevan.

Untuk menjaga keamanan dalam proses pertukaran data, setiap permintaan yang dikirimkan oleh KMS PPM IPB dilengkapi dengan *token request* sebagai mekanisme autentikasi. Apabila *token* yang diberikan valid, maka layanan akan melanjutkan proses pencarian informasi. Sebaliknya, jika token tidak valid, maka permintaan tidak akan diproses lebih lanjut. Mekanisme ini bertujuan untuk memastikan bahwa hanya permintaan yang terverifikasi yang dapat mengakses layanan yang disediakan.

Text Pre Processing

Tahapan *text preprocessing* dalam penelitian ini meliputi *tokenization*, *stop-word removal*, dan *stemming* untuk meningkatkan kualitas data teks sebelum dilakukan analisis lebih lanjut. Proses ini bertujuan untuk mengurangi *noise* serta menormalkan representasi teks sehingga dapat meningkatkan kinerja model dalam mengekstraksi informasi yang relevan (Jurafsky dan Martin 2023). Seluruh proses *text preprocessing* pada penelitian ini diilustrasikan pada Gambar 3.

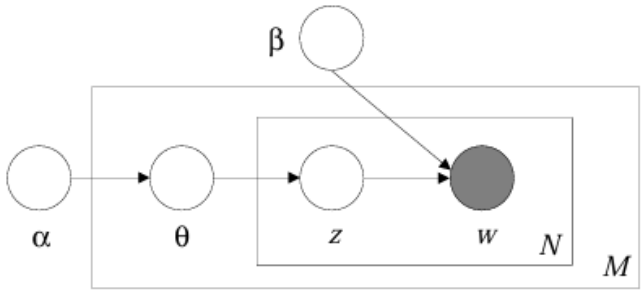


Gambar 3 Alur *Text-preprocessing*

Pengembangan *Latent Dirichlet Allocation*

Latent Dirichlet Allocation (LDA) merupakan model probabilistik generatif yang merepresentasikan dokumen sebagai campuran dari beberapa topik laten, di mana setiap topik didefinisikan sebagai distribusi probabilitas atas kata-kata tertentu (Blei *et al*. 2003).

Seperti yang ditunjukkan pada Gambar 4, struktur LDA divisualisasikan menggunakan konsep *plate notation*, di mana kotak-kotak (*plate*) merepresentasikan proses pengulangan dalam model (Blei *et al*. 2003). *Plate* bagian luar menggambarkan kumpulan dokumen, sedangkan *plate* bagian dalam merepresentasikan distribusi topik dan kata-kata yang berulang dalam setiap dokumen. Pendekatan ini menunjukkan hubungan antara dokumen, topik, dan kata dalam proses generatif, serta menjadi dasar pemahaman dalam model *topic modeling* berbasis probabilistik (Karmila dan Ardianti 2022; Blei *et al.* 2003).



Gambar 4 Representasi Model Grafis LDA

Dalam penelitian ini, model LDA diimplementasikan menggunakan bahasa pemrograman Python. Dari sisi implementasi, pemodelan topik dilakukan dengan menentukan dua topik utama untuk merepresentasikan abstrak PPM yang memiliki panjang teks relatif terbatas. Pemilihan jumlah topik yang terlalu besar berpotensi menyebabkan distribusi topik menjadi terlalu luas dan menghasilkan *noise* pada representasi topik. Berdasarkan pengujian awal, penggunaan dua topik menghasilkan distribusi topik yang lebih stabil dalam merepresentasikan konteks abstrak penelitian. Proses pemodelan dilakukan melalui beberapa iterasi untuk memperoleh distribusi topik yang lebih konsisten. Selanjutnya, distribusi probabilitas topik yang dihasilkan digunakan sebagai representasi teks sebelum dilakukan pengukuran *semantic similarity*.

Pengembangan IR *System*

IR *System* diimplementasikan sebagai bagian dari layanan (*service*) yang dikembangkan menggunakan bahasa pemrograman Python. Setelah distribusi topik dihasilkan melalui pemodelan LDA, IR *System* mengambil alih proses dengan memanfaatkan distribusi topik tersebut dan membandingkannya dengan kata kunci SDGs yang diperoleh dari basis data KMS PPM IPB.

Proses perbandingan dilakukan menggunakan algoritma *semantic similarity* dengan pendekatan *triplet loss*. *Triplet loss* merupakan salah satu metode dalam *metric learning* yang bertujuan untuk mempelajari representasi *embedding* dengan meminimalkan jarak antara pasangan data yang serupa (*anchor dan positive*) serta memaksimalkan jarak terhadap data yang tidak serupa (*negative*) (Schroff *et al.* 2015). Studi terbaru menunjukkan bahwa *triplet loss* efektif dalam meningkatkan kualitas representasi *embedding* untuk berbagai tugas pemrosesan

bahasa alami, termasuk pengukuran *semantic similarity* antar teks (Reimers dan Gurevych 2019; Gao *et al.* 2021). Formulasi yang digunakan dalam *triplet loss* ditunjukkan pada Persamaan 1.

$$\max(||sa - sp|| - ||sa - sn|| + \varepsilon, 0) \quad (1)$$

dengan *sa*, *sp*, dan *sn* masing-masing merepresentasikan *embedding* kalimat untuk *anchor*, *positive*, dan *negative*. Notasi $\| \cdot \|$ menunjukkan metrik jarak yang digunakan, sedangkan parameter margin (ε) memastikan bahwa jarak antara *anchor* dan *positive* lebih kecil dibandingkan dengan jarak antara *anchor* dan *negative* dengan selisih tertentu.

Implementasi *semantic similarity* dilakukan menggunakan pendekatan berbasis *sentence embedding* dengan model *Sentence Transformer paraphrase-multilingual-MiniLM-L12-v2*. Model ini dipilih karena memiliki kemampuan representasi semantik multibahasa, sehingga dapat menangani kemiripan makna pada teks dengan variasi istilah dan bahasa yang berbeda.

Pada penelitian ini, distribusi topik hasil *Latent Dirichlet Allocation* (LDA) digunakan sebagai representasi teks sebelum dilakukan proses *embedding*. Selanjutnya, kata kunci dari masing-masing kategori SDGs pada *master data SDGs* juga diubah menjadi representasi vektor menggunakan model *embedding* yang sama. Proses pengukuran kesamaan dilakukan menggunakan *cosine similarity* antara vektor distribusi topik dokumen PPM dan vektor kata kunci SDGs. Nilai *similarity* dihitung menggunakan pendekatan *cosine distance*, di mana nilai yang semakin mendekati "1" menunjukkan tingkat kedekatan semantik yang semakin tinggi. Model *paraphrase-multilingual-MiniLM-L12-v2* merupakan model *Sentence Transformer* yang dikembangkan menggunakan pendekatan *metric learning* berbasis *triplet loss*, sehingga mampu merepresentasikan hubungan semantik antar teks dengan mempertahankan kedekatan antara data yang relevan dan menjauhkan data yang tidak relevan dalam ruang *embedding*.

Pengujian Integrasi Sistem

Salah satu pengujian sistem difokuskan pada pengujian integrasi antara KMS PPM IPB dan layanan (*service*) IR *System*. Pengujian integrasi merupakan teknik pengujian perangkat lunak di mana modul-modul yang sebelumnya berdiri sendiri digabungkan dan diuji sebagai satu kesatuan sistem (Afifah *et al*. 2024).

Seiring dengan penggunaan metode integrasi berbasis permintaan *Representational State Transfer Application Programming Interface* (REST API), pengujian dilakukan menggunakan Postman. Postman merupakan alat yang digunakan untuk menguji fungsionalitas *Application Programming Interface* (API) melalui pengaturan permintaan (*request*) dan koleksi (*collection*) (Postman, n.d.). Melalui Postman, permintaan API dapat dikirim dalam urutan tertentu untuk mengevaluasi operasi yang kompleks serta memantau aliran data ke dan dari berbagai *endpoint* (Postman, n.d.).

Dalam literatur kualitas perangkat lunak modern, waktu *respons* merupakan bagian dari aspek *performance efficiency* yang berpengaruh terhadap kenyamanan dan persepsi pengguna terhadap sistem (ISO/IEC 25010:2023). Secara umum, *respons* yang mendekati sub-detik cenderung dipersepsikan sebagai interaksi yang cepat dan lancar, sedangkan peningkatan waktu tunggu dapat menurunkan tingkat kepuasan dan kenyamanan pengguna. Temuan ini juga sejalan dengan studi terkait performa sistem terdistribusi modern yang menunjukkan bahwa latensi sistem memiliki dampak langsung terhadap efektivitas interaksi pengguna dengan aplikasi, karena keterlambatan eksekusi akan memengaruhi pengalaman penggunaan secara keseluruhan (Li *et al.* 2021).

Pengujian IR *System* dengan *Precision and Recall* dan *F-Measure*

Pengujian IR *System* difokuskan pada tingkat relevansi dalam proses pencarian data SDGs dengan memanfaatkan metrik *precision* dan *recall*. Evaluasi ini digunakan untuk mengukur sejauh mana sistem mampu mengembalikan dokumen yang relevan terhadap

kebutuhan informasi pengguna. Dalam konteks sistem temu kembali informasi, relevansi tidak hanya ditentukan oleh kesesuaian kata kunci, tetapi oleh kemampuan dokumen dalam memenuhi kebutuhan informasi yang diinginkan pengguna (Sadeli dan Lawanda 2023). Perhitungan nilai *precision* (P) dan *recall* (R) dinyatakan melalui Persamaan 2 dan 3.

$$P = \frac{|TP|}{|TP|+|FP|} \tag{2}$$

$$R = \frac{|TP|}{|TP|+|FN|} \tag{3}$$

Nilai *True Positive* (TP) merepresentasikan jumlah temuan yang relevan, sedangkan *False Positive* (FP) menunjukkan jumlah temuan yang tidak relevan. Adapun *False Negative* (FN) merepresentasikan jumlah dokumen relevan yang tidak berhasil ditemukan oleh sistem.

Selanjutnya, hasil pengukuran tersebut dapat dianalisis menggunakan *F-Measure* atau *F-Score*, yang digunakan untuk mengevaluasi efektivitas sistem penelusuran informasi (Sadeli dan Lawanda 2023). *F-Measure* mengintegrasikan nilai *precision and recall* ke dalam satu metrik yang merepresentasikan keseimbangan antara keduanya. Secara matematis, *F-Measure* dihitung sebagai rata-rata harmonis dari *precision and recall* yang dinyatakan pada Persamaan 4.

$$F1 = 2 \times \frac{P \times R}{R + P} \tag{4}$$

HASIL DAN PEMBAHASAN

Hasil *text-preprocessing*

Tabel 1 menyajikan hasil proses penghapusan kata umum (*stop-word removal*) dan *stemming* pada dokumen abstrak pada data PPM. Hasil tersebut menunjukkan bahwa kata-kata umum serta imbuhan berupa awalan dan akhiran pada dokumen telah berhasil dihilangkan.

Tabel 1 Penghapusan kata-kata *stop-words* dan hasil *stemming*

Dokumen Abstrak	Hasil *Stemming*
Seleksi pejantan unggul dapat dinilai berdasarkan sifat genetik dan penilaian performa kinerja reproduksi dengan tingkat fertilitas. Penilaian fertilitas pejantan secara konvensional dilakukan dengan penilaian Breeding Soundness Examination (BSE), yang dapat dijadikan sebagai parameter tunggul	seleksi jantan unggul nilai dasar sifat genetik nilai performa kerja reproduksi tingkat fertilitas nilai fertilitas jantan cara konvensional laku nilai breeding soundness examination bse jadi bagai parameter tunggul

Hasil Analisis Topik

Hasil analisis topik diperoleh melalui pemodelan LDA. Tabel 2 menunjukkan skor koherensi yang dihasilkan dari kumpulan topik berdasarkan hasil proses *tokenization*. Skor koherensi digunakan untuk mengukur tingkat keterkaitan antar kata dalam suatu topik, dengan rentang nilai antara 0 hingga 1; semakin tinggi nilai yang diperoleh, semakin baik kualitas topik yang dihasilkan.

Kumpulan topik yang telah dihasilkan kemudian dimanfaatkan sebagai representasi semantik dokumen untuk mengevaluasi tingkat kesesuaian topik terhadap SDGs. Proses evaluasi ini dilakukan menggunakan algoritma *semantic similarity*, sehingga keterkaitan antara topik yang dihasilkan dengan kategori SDGs dapat diidentifikasi secara lebih akurat.

Tabel 2 Penghapusan *stop word* dan hasil *stemming*

Distribusi Topik	Hasil Koherensi
prediabetes	0.294
rendah	0.274
ssb	0.225
atur	0.225
tingkat	0.216
jadi	0.190
pengaruh	0.170
baik	0.157
diabetes	0.147
tanda	0.118

Hasil Pencarian Kesesuaian Topik

Langkah selanjutnya adalah melengkapi data yang belum memiliki kategori menggunakan algoritma *semantic similarity*. Pendekatan *triplet loss* diterapkan untuk menentukan nilai relevansi antara sepuluh topik yang dihasilkan dari pemodelan LDA (Tabel 2) dan kata kunci yang terkait dengan masing-masing kategori SDGs. Tabel 3 menyajikan hasil analisis *semantic similarity*. Hasil pengukuran dinyatakan dalam skala relevansi antara 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan tingkat kesamaan yang lebih besar. Berdasarkan hasil tersebut, kategori *Good Health and Well-Being* memperoleh skor tertinggi sebesar 0,230, diikuti oleh *Zero Hunger* dengan skor 0,190, serta *No Poverty* dengan skor 0,169.

Tabel 3 Skor *Semantic Similarity*

#	SDGs No	*SDGs Topic*	Skor Kesamaan
1	3	*Good health and Well-Being*	0.230
2	2	*Zero Hunger*	0.190
3	1	*No Poverty*	0.169
4	7	*Affordable and Clean Energy*	0.133
5	10	*Reduced Inequalities*	0.090
6	5	*Gender Equality*	0.082
7	8	*Decent Work and Economic Growth*	0.081
8	4	*Quality Education*	0.073
9	6	*Clean Water and Sanitation*	0.039
10	12	*Responsible Consumption and Production*	0.035
11	13	*Life on Land*	0.017
12	9	*Industry, Innovation, and Infrastructure*	0.012
13	11	*Sustainable Cities and Communities*	0.009
14	14	*Partnership for the Goals*	0.000

Dominasi kategori *Good Health and Well-Being* mengindikasikan bahwa sebagian besar topik yang dihasilkan oleh model LDA memiliki keterkaitan yang kuat dengan isu kesehatan, seperti peningkatan kualitas hidup, layanan kesehatan, atau kesejahteraan masyarakat. Sementara itu, kemunculan kategori *Zero Hunger* dan *No Poverty* pada peringkat berikutnya menunjukkan adanya keterkaitan topik dengan aspek ketahanan pangan dan kondisi sosial-ekonomi masyarakat. Setiap data PPM dapat dikaitkan dengan maksimal tiga kategori SDGs. Oleh karena itu, tiga kategori dengan skor tertinggi dipilih untuk melengkapi informasi SDGs pada data PPM yang sebelumnya belum memiliki kategori.

Hasil Pengujian Integrasi Sistem

Tahap ini menyajikan hasil pengujian integrasi antara KMS PPM IPB dan layanan (*service*) IR *System* pada kondisi ketika terdapat data PPM yang belum memiliki informasi kategori SDGs. Pengujian dilakukan dengan mengirimkan permintaan (*request*) dari KMS PPM IPB ke *service* IR *System* sebanyak 10 kali untuk mengukur kinerja sistem dalam menangani proses integrasi.

Berdasarkan hasil pengujian pada Tabel 4, diperoleh rata-rata waktu *respons* sebesar 5,48 detik. Menurut Nielsen (1993), waktu *respons* kurang dari 1 detik mampu mempertahankan alur interaksi pengguna secara optimal, sedangkan waktu *respons* hingga 10 detik masih dapat ditoleransi meskipun mulai memengaruhi tingkat kenyamanan pengguna. Temuan ini juga

sejalan dengan teori *Human-Computer Interaction* yang menyatakan bahwa performa sistem memiliki pengaruh signifikan terhadap efektivitas interaksi antara pengguna dan sistem.

Tabel 4 Hasil Pengujian Integrasi

Nomor Tes	Waktu *Respons*
1	5,50
2	5,82
3	5,44
4	5,44
5	5,79
6	5,44
7	5,32
8	5,27
9	5,24
10	5,62

Dengan demikian, waktu respons sebesar 5,48 detik pada sistem yang dikembangkan masih berada dalam batas yang dapat diterima, namun belum mencapai kategori optimal.

Hasil Pengujian IR *System* dengan *Precision and Recall* serta *F-Measure*

Langkah ini difokuskan pada IR *System* menggunakan metrik *precision and recall*. Pengujian dilakukan dengan membandingkan hasil pencarian informasi SDGs yang dihasilkan oleh IR *System* dengan data SDGs yang telah tersedia pada set data PPM.

Evaluasi dilakukan terhadap 20 seperti yang ditunjukan pada Tabel 5. Berdasarkan hasil pengujian terhadap 20 data set pada Tabel 6, diperoleh nilai rata-rata *precision* sebesar 0,233 (23,3%) dan *recall* sebesar 0,375 (37,5%). Nilai tersebut kemudian diintegrasikan menggunakan *F-measure* (*F1-score*) untuk mengevaluasi keseimbangan antara ketepatan dan kelengkapan hasil pencarian, sehingga diperoleh nilai *F1-score* sebesar 0,288 (28,8%).

Tabel 5 *Dataset* yang digunakan untuk pengujian

#	Abstrak	SDGs_NO
1	Salah satu poin yang tertuang dalam garis besar RIP IPB…	2
2	Salah satu permasalahan yang sering dihadapi…	14
3	Sapi potong yang dipeliharanya dapat…	3
4	Penyakit Avian Influenza (AI) adalah penyakit unggas…	3
5	Danau Siombak merupakan danau pasut…	2, 14
6	Cendawan patogen yang menyerang…	15
7	Serangga merupakan komponen…	15
8	Pupuk organik hewan adalah pupuk yang…	3, 8, 4
9	Sprayer merek DOT.ON tipe DTO-16 adalah...	2, 2, 2
10	Tanaman kelor (Moringa oleifera L.) merupakan…	3, 15, 9
11	Kuliah Kerja Nyata – Tematik (KKN-T) Institut Pertanian Bogor...	3, 1, 2
12	Desa Batu Beriga merupakan desa pesisir…	4, 8, 14
13	Desa Bojong Jengkol merupakan salah satu desa…	3
14	Kuliah Kerja Nyata Tematik (KKN-T) IPB University…	3, 4, 17
15	Kondisi pandemi menyebabkan kegiatan...	3, 1, 12
16	The testing was carried based on…	9, 8, 11
17	Kopi adalah salah satu komoditas unggulan ekspor…	9, 8, 12
18	Kentang untuk keripik mempunyai nilai ekonomi…	2, 1, 17
19	Sejumlah dosen IPB University dari...	2, 1, 8
20	Pencemaran dalam sedimen laut…	14, 9, 17

Tabel 6 Hasil *Precision and Recall*

#	Relevan (TP)	Tidak relevan (FP)	Tidak ditemukan (FN)	Recall (R)	Precision (P)
1	1	2	2	0,333333333	0,3333333
2	1	2	0	1	0,3333333
3	1	2	0	1	0,3333333
4	0	3	1	0	0
5	1	2	0	1	0,3333333
6	0	3	1	0	0
7	1	2	2	0,333333333	0,3333333
8	0	3	3	0	0
9	0	3	1	0	0
10	1	2	2	0,333333333	0,3333333
11	1	2	0	1	0,3333333
12	0	3	3	0	0
13	0	3	1	0	0
14	2	1	1	0,666666667	0,6666667
15	1	2	2	0,333333333	0,3333333
16	1	2	2	0,333333333	0,3333333
17	1	2	2	0,333333333	0,3333333
18	0	3	3	0	0
19	1	2	2	0,333333333	0,3333333
20	1	2	1	0,5	0,3333333
Rata - rata *precision dan recall*				0,375	0,23333333
F-Measure				0,287671233	

Nilai *recall* yang lebih tinggi dibandingkan *precision* menunjukkan bahwa IR *System* mampu mengidentifikasi sebagian kategori SDGs yang relevan dari dokumen PPM. Pendekatan berbasis *semantic similarity* dan *sentence embedding* memungkinkan sistem mengenali hubungan makna antar teks meskipun tidak memiliki kesamaan leksikal secara langsung. Hal ini menunjukkan bahwa representasi semantik yang digunakan cukup efektif dalam memperluas cakupan pencarian informasi. Namun demikian, nilai *precision* yang relatif rendah menunjukkan bahwa sistem masih menghasilkan cukup banyak *false positive*. Berdasarkan hasil pengujian, sebagian besar dokumen menghasilkan pola pemetaan dengan satu kategori relevan yang diikuti oleh dua hingga tiga kategori yang tidak relevan. Kondisi ini mengindikasikan bahwa pendekatan *semantic similarity* yang digunakan cenderung bersifat permisif dalam menentukan kedekatan semantik antar kategori SDGs. Selain itu, karakteristik kategori SDGs yang saling beririsan secara konseptual (*overlapping topics*) turut memengaruhi hasil pemetaan. Kondisi tersebut menyebabkan sistem masih mengalami kesulitan dalam membedakan relevansi kontekstual secara lebih spesifik. Penggunaan distribusi topik hasil *Latent Dirichlet Allocation* (LDA) sebagai representasi juga memengaruhi hasil pengukuran *semantic similarity*. Pada beberapa abstrak dengan cakupan topik yang luas, distribusi topik yang dihasilkan masih bersifat umum sehingga representasi semantik antar kategori SDGs menjadi saling beririsan. Akibatnya, sistem cenderung menghasilkan pemetaan kategori yang lebih luas dibandingkan label *ground truth* yang ditentukan secara manual. Hasil pengujian juga menunjukkan bahwa mekanisme penentuan *threshold similarity* masih memerlukan optimalisasi. Nilai *threshold* yang terlalu rendah berpotensi meningkatkan jumlah kategori yang teridentifikasi sebagai relevan, namun berdampak pada meningkatnya *false positive* dan penurunan *precision*. Oleh karena itu, pengembangan lebih lanjut diperlukan, khususnya dalam optimalisasi representasi topik, pengaturan *threshold similarity* yang lebih adaptif, serta pengayaan *master data* SDGs untuk meningkatkan ketepatan hasil pemetaan informasi.

SIMPULAN

Hasil penelitian menunjukkan bahwa pendekatan berbasis *Information Retrieval System* yang dikombinasikan dengan pemodelan topik *Latent Dirichlet Allocation* (LDA) dan *semantic*

similarity berbasis *sentence embedding* dapat digunakan untuk membantu proses otomatisasi pelabelan *Sustainable Development Goals* (SDGs) pada data penelitian dan pengabdian kepada masyarakat (PPM) di IPB University. Hasil evaluasi menunjukkan nilai *F1-score* sebesar 28,8%, yang mengindikasikan bahwa sistem telah mampu mengidentifikasi sebagian kategori SDGs yang relevan, meskipun ketepatan hasil pemetaan masih perlu ditingkatkan. Hali ini dikareanakan pendekatan masih berfokus pada kesamaan semantik berbasis representasi kata dan topik, bukan pemahaman konteks dokumen secara menyeluruh. Selain itu, pengujian integrasi antara KMS PPM IPB dan layanan *Information Retrieval System* menunjukkan rata-rata waktu respons sebesar 5,48 detik, yang masih berada dalam batas yang dapat diterima untuk sistem terintegrasi.

UCAPAN TERIMA KASIH

Penulis menyampaikan rasa terima kasih kepada Lembaga Manajemen Informasi dan Transformasi Digital (LMITD) serta Direktorat Riset dan Inovasi (DRI) Institut Pertanian Bogor atas dukungan, fasilitas, dan informasi yang telah diberikan selama proses penelitian ini berlangsung. Penulis mengakui penggunaan teknologi kecerdasan buatan dalam membantu memperbaiki struktur penulisan, tata bahasa, serta memberikan saran tambahan dalam penyusunan naskah penelitian ini. Seluruh isi, analisis, interpretasi data, dan kesimpulan yang disajikan dalam penelitian ini sepenuhnya merupakan tanggung jawab penulis.

DAFTAR PUSTAKA

Afifah HR, Kusumo DS, Selviandro N. 2024. Pengimplementasian integration testing, performance testing, dan user acceptance testing pada aplikasi Cafeasy berbasis web (Studi kasus: Café Daerah Bandung). *E-Proceeding of Engineering* [Internet], [diunduh 2026 Jan 13]; 11(4):5151-5159.

Adiyanto AT, Handayani D. 2022. Information retrieval sistem kearsipan pencarian dokumen di Dinas Pemberdayaan Perempuan dan Perlindungan Anak Kota Semarang menggunakan metode vector space model. *Jurnal Mahajana Informasi* [Internet], [diunduh 2026 Apr 19]; 7(1):2527-8290. https://doi.org/10.51544/jurnalmi.v7i1.2538

Blei DM, Andrew Y, Jordan M. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* [Internet]. [diunduh 2026 Jan 13]; 3:993-1022. https://dx.doi.org/http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

Gao T, Yao X, Chen D. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [Internet]. [diunduh 2026 April 19]; p. 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552.

Guisiano JE, Chiky R, De Mello J. 2022. SDG-Meter: A deep learning based tool for automatic text classification of the sustainable development goals. In: *Lecture Notes in Computer Science* [Internet]. [diunduh 2026 Feb 21]; p. 259-271. https://doi.org/10.1007/978-3-031-21743-2_21

Hajikhani A, Souminen. 2022. Mapping the sustainable development goals (SDGs) in science, technology and innovation: application of machine learning in SDG-oriented artefact detection. *Sensors (Basel)* [Internet]. [diunduh 2026 April 19]; 127: 6661–6693. https://doi.org/10.1007/s11192-022-04358-x.

Hsu DF, LaFleur M, Orazbek I. 2022. Improving SDG classification precision using combinatorial fusion. *An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy* [Internet]. [diunduh 2026 April 19]; 22(3):1067. https://doi.org/10.3390/s22031067.

Jurafsky D, Martin JH. 2023. *Speech and Language Processing*. Stanford (US): Stanford University.

Kalmukov Y. 2022. A comparison between Latent Semantic Analysis and Vector Space Model for document similarity. *International Journal of Advanced Computer Science and*

Applications (IJACSA) [Internet]. [diunduh 2026 April 19]; 13(2): 74–80. http://dx.doi.org/10.14569/IJACSA.2022.0130209

Karmila S, Ardianti VI. 2022. Metode Latent Dirichlet Allocation untuk menentukan topik teks suatu berita. *Jurnal Informatika & Komputasi* [Internet]. [diunduh 2026 Feb 1]; 16(01): 36-44. https://doi.org/10.56956/jiki.v16i01.100.

Lin J, Nogueira R, Yates A. 2022. Pretrained transformers for text ranking: BERT and beyond. In: *Synthetic Lectures on Human Language Technologies* [Internet]. [diunduh 2026 Apr 19]; https://doi.org/10.1007/978-3-031-02181-7

Li N, Tao LV, Wang X, Meng X, Xu J, Guo Y. 2025. Research progress and hot topics of distributed photovoltaic: Bibliometric analysis and Latent Dirichlet Allocation model. *Energy and Buildings* [Internet]. [diunduh 2026 Apr 19]; 327. https://doi.org/10.1016/j.enbuild.2024.115056.

Li S, Zhang H, Jia Z, Zhong C, Zhang C, Shan Z, Shen J, Babar MA. 2021. Understanding and addressing quality attributes of microservices architecture: A systematic literature review. Information and Software Technology [Internet]. [diunduh 2026 April 19]; 131(3-4):106449. https://doi.org/10.1016/j.infsof.2020.106449

Makmum WW, Ningrum IP, Sajiah AM. 2022. Pretrained transformers for text ranking: BERT and beyond. *Jurnal semanTIK* [Internet]. [diunduh 2026 Feb 1]; 8(1): 69-76 https://doi.org/10.55679/semantik.v8i1.15346.

Nielsen J. 1993. *Usability Engineering*. Boston (US): Academic Press.

Pal A, Mukhopadhyay P. 2024. Categorisation of Indian Research Publications by sustainable development goals (SDGs): A machine learning approach. *Journal of Information and Knowledge* [Internet]. [diunduh 2026 Apr 19]; 61(6):303-311 https://doi.org/10.17821/srels/2024/v61i6/171637.

Reimers N, Gurevych I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [Internet]. [diunduh 2026 April 19]; p. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

Rinaldi FM, Giuffrida G, Nicotra S, Dispinseri F. 2024. *A Classification Algorithm to Link Official Documents to Sustainable Development Goals*. London (UK): Taylor & Francis.

Rulandari N. 2021. Study of sustainable development goals (SDGS) quality education in indonesia in the first three years. *Budapest International Research and Critics Institute (BIRCI-Journal) Humanities and Social Sciences* [Internet]. [diunduh 2026 Feb 21]; 4(2):2702-2708. https://doi.org/10.33258/birci.v4i2.1978.

Sadeli AF, Lawanda II. 2023. Recall, precision, and F-measure for evaluating information retrieval systtem in Electronic Document Management Systems (EDMS). *Khizanah al-Hikmah: Jurnal Ilmu Perpustakaan, Informasi, & Kearsipan*. 11 (2): 231-241. https://doi.org/10.24252/kah.v11i2a8

Schroff F, Kalenichenko D, Philbin J. 2015. FaceNet: A unified embedding for face recognition and clustering. In: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [Internet]. [diunduh 2026 Mar 25]; p. 815-823 https://doi.org/10.1109/CVPR.2015.7298682.

Sharma C, Sharma S, Sakshi. 2022. Latent dirichlet allocation (LDA) based information modelling on Blockchain technology: a review of trends and research patterns used in integration. *Multimedia Tools and Applications* [Internet]. [diunduh 2026 Apr 19]; 81(25): 36805–36831. https://doi.org/10.1007/s11042-022-13500-z