JURNAL PUSTAKA AI
JURNAL PUSAT AKSES KAJIAN TEKNOLOGI ARTIFICIAL INTELLIGENCE
Pustaka Galeri Mandiri
Vol. 6 No. 1 (2026) 265 – 275
E ISSN: 2809-4069

Progressive Topic Formation: *Framework Iteratif* untuk Membangun *Domain Knowledge Collection*

Maulana Aziz Assuja[1], Saniati[2*], Yeni Agus Nurhuda[3]
[1,2] Sistem dan Teknologi Informasi, Institut Teknologi dan Bisnis Mesuji,
[3]Teknologi Informasi, Institut Bisnis dan Teknologi Kalimantan,
[1]aziz.maulana.assuja@gmail.com [2]saniati.8820@gmail.com [3]nurhuda5875@gmail.com

Abstract

Topic modeling is widely used to identify topic structures in large document collections but its output generally focuses on topic representations and does not directly produce reusable domain knowledge. This study proposes Progressive Topic Formation (PTF), a framework for constructing Domain Knowledge Collection (DKC) through Semantic Domain Prototype, semantic retrieval, progressive document clustering and Domain Knowledge Construction. The framework was applied to 599,059 Indonesian Wikipedia articles, with the 1,000 highest-scoring documents in the Education domain used for clustering. The process produced 173 clusters without noise, with a mean semantic coherence of 0.8573. Cluster size showed a strong negative correlation with coherence. The resulting DKC consists of topic names, summaries, keywords, topic representations and supporting documents. Topic representations showed high similarity to keywords (0.8192) and summaries (0.7921) while Representation–Document Similarity was 0.6626 and negatively correlated with cluster size. These results demonstrate that PTF can transform semantically related document groups into structured and reusable domain knowledge for Knowledge Management Systems, semantic search and Retrieval-Augmented Generation (RAG).

Keywords: Progressive Topic Formation; Semantic Retrieval; Progressive Document Clustering; Domain Knowledge Collection; Large Language Model.

Abstrak

Topic modeling banyak digunakan untuk mengidentifikasi struktur topik dari koleksi dokumen berskala besar, tetapi keluarannya umumnya berfokus pada representasi topik dan belum secara langsung membentuk pengetahuan domain yang dapat digunakan kembali. Penelitian ini mengusulkan Progressive Topic Formation (PTF), framework untuk membangun Domain Knowledge Collection (DKC) melalui Semantic Domain Prototype, semantic retrieval, progressive document clustering, dan Domain Knowledge Construction. Framework diterapkan pada 599.059 artikel Wikipedia Bahasa Indonesia, dengan 1.000 dokumen berskor tertinggi pada domain Pendidikan digunakan untuk proses clustering. Proses tersebut menghasilkan 173 cluster tanpa noise dengan rata-rata semantic coherence 0,8573. Ukuran cluster memiliki korelasi negatif yang sangat kuat terhadap coherence. DKC yang dihasilkan terdiri atas nama topik, ringkasan, kata kunci, representasi topik, dan dokumen pendukung. Representasi topik memiliki similarity tinggi terhadap kata kunci (0,8192) dan ringkasan (0,7921), sedangkan Representation–Document Similarity sebesar 0,6626 dan berkorelasi negatif terhadap ukuran cluster. Hasil tersebut menunjukkan bahwa PTF mampu mentransformasikan kelompok dokumen yang memiliki keterkaitan semantik menjadi pengetahuan domain yang terstruktur dan dapat digunakan kembali pada Knowledge Management System, semantic search, dan Retrieval-Augmented Generation (RAG).

Kata kunci: *Progressive Topic Formation; Semantic Retrieval; Progressive Document Clustering; Domain Knowledge Collection; Large Language Model.*

Submitted : 13-06-2026 | Reviewed : 20-06-2026 | Accepted : 27-06-2026

1. Pendahuluan

Transformasi digital menghasilkan koleksi dokumen dalam jumlah besar pada berbagai sektor, sehingga diperlukan metode yang mampu mengorganisasi dan mengekstraksi informasi dari data tidak terstruktur. Perkembangan *semantic embedding, transformer, dan Retrieval-Augmented Generation* (RAG) telah meningkatkan kemampuan sistem dalam memahami hubungan semantik antar dokumen untuk mendukung *semantic search* dan *question answering,* sebagaimana ditunjukkan oleh Lewis et al. [1] melalui penggabungan pencarian informasi dan generasi teks, serta Gao et al.[2] melalui kajiannya mengenai pengembangan RAG untuk menghubungkan LLM dengan sumber informasi eksternal. Meskipun demikian, sebagian besar pendekatan masih berfokus pada *retrieval* informasi dan belum menghasilkan representasi pengetahuan yang dapat dimanfaatkan kembali.

Topic modeling merupakan pendekatan yang banyak digunakan untuk mengidentifikasi tema utama dalam koleksi dokumen. LDA yang diperkenalkan oleh Blei et al.[3] memodelkan topik berdasarkan distribusi probabilitas kata, sedangkan BERTopic yang dikembangkan oleh Grootendorst [4] memanfaatkan *semantic embedding* dalam proses pembentukan topik. Namun, keluaran kedua metode tersebut umumnya terbatas pada label atau kumpulan kata representatif sehingga belum membentuk *Domain Knowledge Collection* (DKC) yang mampu merepresentasikan pengetahuan secara lebih komprehensif.

Kemajuan *Large Language Model* (LLM) dan *semantic retrieval* membuka peluang untuk mengekstraksi konsep dan mensintesis informasi secara lebih semantik. Namun, pemanfaatannya dalam konteks tersebut masih berorientasi pada proses inferensi, sehingga pengetahuan yang dihasilkan belum berkembang menjadi koleksi pengetahuan yang dapat diperbarui dan digunakan kembali. Kondisi tersebut menunjukkan perlunya mekanisme *Progressive Topic Formation* (PTF), yaitu proses pembentukan pengetahuan yang berlangsung secara bertahap melalui integrasi *semantic retrieval* dan LLM.

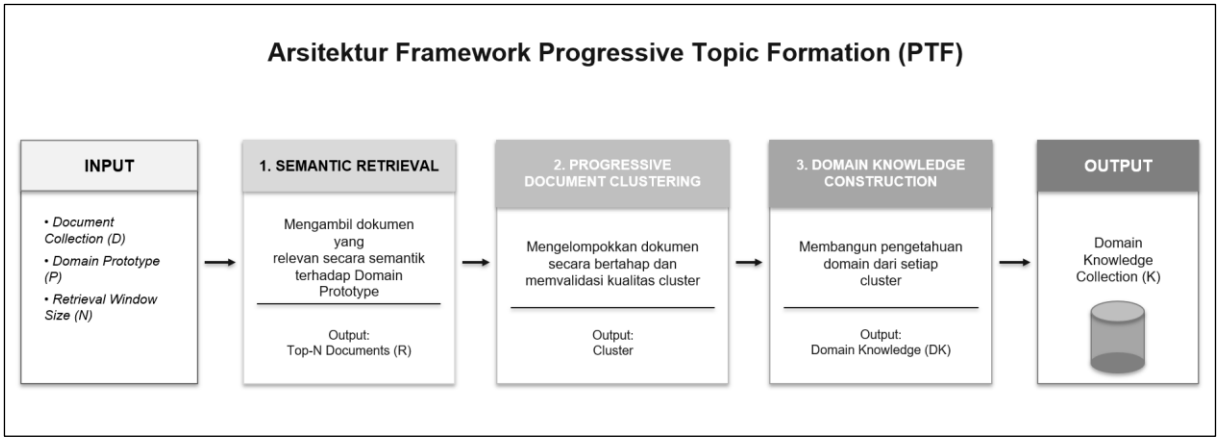
Penelitian ini mengusulkan PTF sebagai *framework* untuk membangun DKC dari koleksi dokumen tidak terstruktur. *Framework* memperkenalkan *Semantic Domain Prototype* sebagai representasi awal domain yang diproses melalui tahapan *semantic retrieval, progressive document clustering,* dan *Domain Knowledge Construction* untuk mentransformasikan *cluster* dokumen menjadi *knowledge object* yang memiliki representasi semantik terstruktur. Kontribusi utama penelitian ini adalah menggeser keluaran *topic modeling* dari sekadar representasi topik menjadi DKC yang dapat digunakan kembali sebagai *knowledge layer*.

2. Metode Penelitian

2.1. Gambaran Umum Framework

PTF membangun DKC melalui serangkaian tahapan yang dimulai dari *Semantic Domain Prototype* sebagai representasi awal domain. *Prototype* digunakan pada proses *semantic retrieval* untuk memperoleh dokumen-dokumen yang memiliki keterkaitan semantik terhadap domain yang diinginkan. Dokumen hasil *retrieval* kemudian diproses secara bertahap menggunakan *progressive document clustering* hingga terbentuk *cluster* yang memenuhi kriteria kualitas. Setiap *cluster* yang valid selanjutnya diproses melalui *Domain Knowledge Construction* untuk menghasilkan *Domain Knowledge* yang terdiri atas *Topic Name, Topic Summary, Topic Representation, Topic Keywords,* serta jumlah dokumen pendukung. Representasi topik tersebut kemudian diubah menjadi *embedding* dan disimpan sebagai bagian dari *Domain Knowledge Collection*.

Submitted : 13-06-2026 | Reviewed : 20-06-2026 | Accepted : 27-06-2026



Gambar 1. Arsitektur *Framework* PTF

Arsitektur pada Gambar 1 menunjukkan tahapan utama *framework* PTF secara konseptual, yang meliputi *semantic retrieval, progressive document clustering*, dan *domain knowledge construction*. Implementasi setiap tahapan tersebut secara lebih rinci diformulasikan dalam algoritma PTF sebagaimana disajikan pada Tabel 1.

Tabel 1. Algoritma PTF

Algoritma : Progressive Topic Formation (PTF)
Input: D : Document Collection P : Domain Prototype N : Retrieval Window Size
Output: K : Domain Knowledge Collection
1: R ← SemanticRetrieval(D, P) 2: nextCluster ← 0 3: while Exists(Unclustered(R)) do 4: B ← SelectTopDocuments(R, N) 5: C ← ClusterDocuments(B) 6: for each cluster c ∈ C do 7: if IsValidCluster(c) then 8: AssignClusterID(c, nextCluster) 9: MarkClustered(c) 10: nextCluster ← nextCluster + 1 11: else 12: MarkAsUnclustered(c) 13: end if 14: end for 15: end while 16: for each Cluster c do 17: DK ← BuildDomainKnowledge(c) 18: DK.embedding ← Embed(DK.topic_representation) 19: DK.document_count ← CountDocuments(c) 20: Store(DK)

Submitted : 13-06-2026| Reviewed : 20-06-2026 | Accepted : 30-06-2026

21: end for
22: return K

2.2. Semantic Retrieval

Tahap pertama pada *Progressive Topic Formation* bertujuan memperoleh dokumen yang memiliki keterkaitan semantik terhadap *Domain Prototype*. *Prototype* adalah deskripsi naratif dari sebuah domain. Deskripsi tersebut kemudian direpresentasikan sebagai *embedding* dan digunakan sebagai *query* pada indeks vektor untuk memperoleh dokumen dengan tingkat kemiripan tertinggi. Mekanisme ini mengikuti pendekatan *dense retrieval* yang dikembangkan oleh Karpukhin et al. [5], yang melakukan pencarian berdasarkan kedekatan representasi *query* dan dokumen dalam ruang *vektor*. Berbeda dengan pencarian berbasis kata kunci, *semantic retrieval* memungkinkan sistem menemukan dokumen yang memiliki makna serupa meskipun menggunakan istilah yang berbeda. Hasil *retrieval* kemudian diurutkan berdasarkan nilai kemiripan semantik dan menjadi masukan bagi proses pembentukan topik pada tahap berikutnya.

2.3. Progressive Document Clustering

Dokumen hasil *semantic retrieval* diproses secara bertahap menggunakan *retrieval window* berukuran tetap (N dokumen) dan diurutkan menurun berdasarkan *semantic score*. Pada setiap iterasi, sejumlah N dokumen dipilih untuk membentuk kandidat *cluster*. Pendekatan ini memungkinkan proses *clustering* dilakukan secara progresif tanpa harus mengelompokkan seluruh dokumen hasil *retrieval* secara bersamaan.

Framework tidak mengharuskan penggunaan algoritma *clustering* tertentu, namun perlu diperhatikan karakteristik data *embedding* (*sparse vs dense embedding*) yang dihasilkan dari proses sebelumnya. Berbagai pendekatan telah digunakan untuk mengelompokkan data berdasarkan kemiripan representasi seperti algoritma K-Means oleh Hardi et al. [6], *Hierarchical Clustering* oleh Patidar et al.[7], DBSCAN, HDBSCAN pada penelitian Eklund et al.[8] , juga penerapan beberapa algoritma *cluster* untuk mengetahui perbandingan perfomanya [9]. Namun demikian algoritma seperti K-Means yang memerlukan jumlah *cluster* yang ditentukan di awal, dianggap kurang sesuai dengan karakteristik *framework* pada penelitian ini.

Pada implementasi acuan penelitian ini digunakan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) yang dikembangkan oleh oleh Campello et al. [10], karena mampu mengidentifikasi jumlah *cluster* secara otomatis tanpa menentukan jumlah *cluster* di awal, serta mendeteksi dokumen yang tidak membentuk *cluster* (*noise*) berdasarkan kepadatan data [11]. *Cluster* yang memenuhi kriteria kualitas diberikan *Cluster ID*, sedangkan dokumen yang belum membentuk *cluster* yang valid akan diproses kembali pada iterasi berikutnya hingga seluruh dokumen berhasil dipetakan atau tidak lagi memenuhi syarat pembentukan *cluster*. Untuk menjamin bahwa seluruh dokumen tetap terwakili dalam lapisan pengetahuan, dokumen yang tetap tidak dapat membentuk *cluster* hingga proses berakhir diperlakukan sebagai *singleton cluster*, yaitu satu dokumen merepresentasikan satu *cluster*.

2.4. Domain Knowledge Construction

Setelah *cluster* yang valid diperoleh, setiap *cluster* diperlakukan sebagai representasi awal suatu topik. Seluruh dokumen dalam *cluster* digunakan sebagai masukan untuk membangun *Domain Knowledge* menggunakan LLM. Proses ini menghasilkan empat komponen utama, yaitu nama topik (*topic name*), ringkasan topik (*topic summary*), kata kunci utama (*topic keywords*), dan representasi topik (*topic representation*). Berbeda dengan *topic modeling* konvensional yang menghasilkan distribusi kata, pendekatan ini menghasilkan representasi pengetahuan yang dapat dipahami secara langsung.

2.5. Topic Representation Embedding

Representasi topik (*topic representation*) yang dihasilkan kemudian dikonversi kembali menjadi *embedding* menggunakan model representasi semantik yang sama dengan proses *semantic retrieval*. *Embedding* ini berfungsi sebagai representasi vektor dari setiap *Domain Knowledge* sehingga memungkinkan pencarian berbasis semantik pada tingkat topik. Selain *embedding*, sistem juga mencatat jumlah dokumen pendukung (*document count*) sebagai informasi mengenai tingkat dukungan setiap topik terhadap koleksi dokumen.

Submitted : 13-06-2026| Reviewed : 20-06-2026 | Accepted : 30-06-2026

2.6. Domain Knowledge Collection

Seluruh *Domain Knowledge* yang berhasil dibentuk disimpan ke dalam *Domain Knowledge Collection* (DKC) sebagai keluaran akhir *Progressive Topic Formation*. Setiap entitas dalam DKC terdiri atas nama topik, ringkasan topik, kata kunci utama, representasi topik, *embedding* representasi topik, serta jumlah dokumen pendukung. Struktur tersebut memungkinkan DKC dimanfaatkan sebagai lapisan pengetahuan (*knowledge layer*) pada berbagai aplikasi seperti *Knowledge Management System*, *semantic search*, maupun RAG.

2.7. Dataset Penelitian

Penelitian ini menggunakan 599.059 dokumen dari korpus Wikipedia Bahasa Indonesia versi November 2023 yang tersedia melalui Indonesian NLP pada *Hugging Face*[12]. Wikipedia dipilih karena memiliki cakupan domain yang luas, struktur artikel yang relatif konsisten, serta telah banyak digunakan sebagai korpus referensi pada berbagai penelitian NLP. Seluruh artikel diproses melalui tahapan *adaptive paragraph-aware chunking* untuk menghasilkan unit-unit teks yang lebih sesuai bagi proses *embedding* dan *semantic retrieval*. Setiap *chunk* kemudian direpresentasikan ke dalam *embedding* semantik dan disimpan pada indeks vektor sehingga dapat dilakukan pencarian berdasarkan kemiripan semantik.

2.8. Skenario Pengujian

Pengujian dilakukan untuk menganalisis kemampuan PTF dalam membangun DKC secara bertahap melalui tiga tahap utama, yaitu *semantic retrieval, progressive document clustering,* dan *Domain Knowledge Construction* menggunakan dataset yang telah ditentukan. Pengujian PTF ditujukan untuk menganalisis karakteristik pengetahuan yang terbentuk pada setiap tahap yang mencakup tiga aspek utama, yaitu (1) relevansi dokumen pada tahap *semantic retrieval*, (2) homogenitas, distribusi, dan keterpisahan semantik pada tahap *clustering*, serta (3) kesesuaian dan tingkat abstraksi representasi pengetahuan pada tahap *Domain Knowledge Construction*.

Pada tahap *semantic retrieval*, pengujian difokuskan pada kemampuan memperoleh dokumen yang memiliki keterkaitan semantik dengan *Semantic Domain Prototype* dengan mengambil top K dokumen dengan *semantic similarity* tertinggi. Hasil tahap ini digunakan sebagai masukan bagi proses *progressive document clustering* dan dianalisis berdasarkan karakteristik serta kesesuaiannya dengan domain yang ditentukan.

Pada tahap *progressive document clustering*, pengujian mencakup keberhasilan pembentukan *cluster*, *homogenitas* semantik dalam *cluster*, hubungan ukuran *cluster* terhadap homogenitas, serta kedekatan semantik antar-*cluster*. Keberhasilan pembentukan *cluster* dianalisis berdasarkan jumlah *cluster*, jumlah dokumen yang terkelompok, jumlah *noise*, rasio *noise*, dan ukuran *cluster*. Homogenitas semantik diukur menggunakan *semantic coherence*, yang digunakan untuk menilai keterkaitan semantik antar-elemen dalam suatu topik atau kelompok [13]. Pengukuran meliputi *mean, median, minimum, maximum,* dan *weighted coherence*. Hubungan antara ukuran *cluster* dan *semantic coherence* dianalisis menggunakan korelasi Spearman dan Pearson, sedangkan kedekatan antar-*cluster* dianalisis menggunakan *inter-cluster similarity* sebagai indikator hubungan antarkelompok. Pendekatan ini sejalan dengan evaluasi validitas clustering yang mempertimbangkan *intra-cluster cohesion* dan *inter-cluster separation* sebagai karakteristik penting kualitas pengelompokan [14].

Pada tahap *Domain Knowledge Construction*, pengujian difokuskan pada kesesuaian semantik *topic representation* terhadap informasi yang direpresentasikan. Pengukuran dilakukan menggunakan *Keyword–Representation Similarity, Representation–Summary Similarity,* dan *Representation–Document Similarity*. Pengukuran tersebut menggunakan representasi *embedding* untuk membandingkan kedekatan semantik antar-teks, sebagaimana pendekatan *Sentence*-BERT yang menghasilkan representasi semantik dan memungkinkan perbandingan menggunakan *cosine similarity* [15]. Pemanfaatan *embedding* untuk merepresentasikan dokumen dan membentuk kelompok semantik juga sejalan dengan pendekatan BERTopic, yang menggunakan *document embedding, clustering*, dan kemudian menghasilkan representasi topik dari kelompok dokumen [4]. Masing-masing ukuran dianalisis menggunakan *mean, median*, dan standar deviasi. Selain itu, hubungan antara ukuran *cluster* dan *Representation–Document Similarity* dianalisis menggunakan korelasi Spearman untuk mengetahui kecenderungan perubahan kedekatan representasi terhadap dokumen sumber seiring bertambahnya ukuran *cluster*.

Konfigurasi implementasi yang digunakan selama pengujian ditunjukkan pada Tabel 2. Konfigurasi tersebut merupakan referensi implementasi dan tidak membatasi penggunaan teknologi lain selama mengikuti prinsip kerja *framework* yang diusulkan.

Tabel 2. Konfigurasi Eksperimen

Komponen	Konfigurasi	Parameter Spesifik
Embedding Model	*BGE-M3*	-
Retrieval Engine	*Elasticsearch Vector Search*	Jenis *vector : Dense Vector* *Similarity : Cosine* *Top K : 1000* *Max Candidate: 5 x Top K*
Clustering	*Window Size (N)*	*50*
	HDBSCAN	*Min cluster size : 2* *Min samples : 2* *Cluster selection method : leaf*
Large Language Model	*Qwen 3.5 4B*	*Max Tokens: 128K* *Reasoning: disabled* *Temperature: 0*

3. Hasil dan Pembahasan

3.1. Hasil Semantic Retrieval

Berdasarkan dataset yang ditentukan, dari koleksi awal sebanyak 599.059 dokumen Wikipedia Bahasa Indonesia. Seluruh korpus diproses melalui *semantic retrieval* dengan menggunakan *Semantic Domain Prototype* pada Domain Pendidikan. *Prototype* domain pendidikan memiliki deskripsi sebagai berikut:

"Semua topik mengenai pendidikan dasar, pendidikan menengah, pendidikan tinggi, sekolah, universitas, guru, dosen, mahasiswa, kurikulum, pembelajaran, e-learning, penelitian pendidikan, evaluasi pendidikan, akademik, dan pengembangan kompetensi.".

Deskripsi *Prototype* dikonversi menjadi *embedding* semantik dan digunakan untuk memperoleh dokumen yang memiliki kedekatan makna dengan domain yang ditentukan. Dari tahap ini, kemudian dipilih 1.000 (Top K) dokumen hasil *retrieval* dengan skor *semantic retrieval* tertinggi sebagai masukan untuk *progressive document clustering.*

Hasil *retrieval* menunjukkan bahwa 1.000 dokumen yang diperoleh mencakup berbagai aspek yang berkaitan dengan pendidikan, antara lain pendidikan menengah, perguruan tinggi, kepemimpinan akademik, kurikulum, prestasi pendidikan, dan institusi pendidikan. Hal ini menunjukkan bahwa *Semantic Domain Prototype* dapat digunakan untuk membatasi ruang pencarian berdasarkan kedekatan semantik sebelum proses pembentukan *cluster* dilakukan. Contoh dokumen hasil *semantic retrieval* ditunjukkan pada Tabel 3.

Tabel 3. Contoh Hasil *Domain Retrieval*

Judul Dokumen	Representasi Domain
Sekolah Menengah Atas Berstandar Internasional dan Prestasi Akademik	Pendidikan menengah, kurikulum internasional, prestasi akademik
Struktur Kepemimpinan dan Model Universitas Penelitian	Kepemimpinan akademik, universitas riset
Profil dan Prestasi Sekolah Menengah di Jawa Barat	Pendidikan menengah, sekolah unggulan
Perguruan Tinggi dan Kepemimpinan Akademik di Indonesia	Perguruan tinggi, manajemen akademik
…	….

3.2. Hasil Progressive Document Clustering

Sebanyak 1.000 dokumen hasil *semantic retrieval* dikelompokkan menggunakan HDBSCAN dan menghasilkan 173 *cluster*, tanpa dokumen yang dikategorikan sebagai *noise*. Ukuran *cluster* bervariasi dari 2 hingga 48

dokumen, dengan rata-rata 5,78 dokumen per *cluster*. Pengukuran *semantic coherence* menghasilkan nilai rata-rata 0,8573 dan median 0,8702, dengan rentang 0,6794–0,9545. Nilai *weighted coherence* sebesar 0,7914 menunjukkan bahwa ketika kontribusi setiap *cluster* diperhitungkan berdasarkan jumlah dokumen, koherensi keseluruhan lebih rendah dibandingkan rata-rata antar-*cluster*.

Tabel 4. Statistik Hasil *Progressive Document Clustering*

Parameter	Hasil
Dokumen hasil *retrieval*	1.000
Jumlah *cluster*	173
Dokumen ter-*cluster*	1.000
Noise	0
Rasio *noise*	0,0000
Rata-rata ukuran *cluster*	5,78
Mean coherence	0,8573
Median coherence	0,8702
Weighted coherence	0,7914
Minimum coherence	0,6794
Maximum coherence	0,9545

Distribusi ukuran *cluster* memperlihatkan bahwa sebagian besar cluster berukuran kecil. Berdasarkan pengelompokan ukuran, terdapat 152 *cluster* kecil (≤9 dokumen), 8 *cluster* menengah (10–19 dokumen), dan 13 *cluster* besar (≥20 dokumen). Ketiga kelompok tersebut masing-masing mencakup 456, 106, dan 438 dokumen. Dengan demikian, hanya sekitar 7,5% *cluster* yang termasuk kategori besar, tetapi kelompok tersebut mencakup 43,8% dari seluruh dokumen yang berhasil dikelompokkan.

Analisis lebih lanjut menunjukkan adanya hubungan negatif yang sangat kuat antara ukuran *cluster* dan *semantic coherence*. Nilai korelasi Spearman sebesar $\rho = -0{,}9269$ ($p < 0{,}001$) menunjukkan bahwa semakin besar jumlah dokumen dalam *cluster*, semakin rendah kecenderungan koherensi semantiknya. Hubungan yang sama juga terlihat melalui korelasi Pearson sebesar $r = -0{,}7564$ ($p < 0{,}001$). Dengan demikian, ukuran *cluster* merupakan faktor yang berkaitan kuat dengan homogenitas semantik *cluster*. Temuan tersebut memberikan interpretasi penting terhadap hasil *clustering*. *Cluster* kecil cenderung memiliki koherensi yang tinggi, sedangkan *cluster* besar memiliki koherensi yang relatif lebih rendah.

Selain *within-cluster coherence*, evaluasi dilakukan terhadap kedekatan antar-*cluster*. Rata-rata *inter-cluster similarity* sebesar 0,6458 dengan *median* 0,6384 menunjukkan bahwa *cluster* yang terbentuk tidak sepenuhnya terpisah secara semantik. Nilai *similarity* tertinggi mencapai 0,9804, terutama terjadi pada pasangan *cluster* berukuran besar. Kondisi tersebut menunjukkan bahwa beberapa *cluster* besar memiliki kedekatan semantik yang tinggi satu sama lain dan berpotensi merepresentasikan wilayah konsep yang beririsan.

Temuan ini penting karena menunjukkan bahwa keberhasilan clustering tidak cukup dinilai hanya berdasarkan jumlah *cluster* atau nilai *coherence* rata-rata. *Cluster* besar dapat mencakup wilayah semantik yang lebih luas dan berpotensi mengalami *semantic drift*. Dengan demikian, hasil *clustering* perlu dipertimbangkan kembali pada tahap pembentukan DKC agar abstraksi yang dihasilkan tidak kehilangan struktur konseptual domain.

Submitted : 13-06-2026| Reviewed : 20-06-2026 | Accepted : 30-06-2026

3.3. Hasil Domain Knowledge Construction

Cluster yang terbentuk selanjutnya ditransformasikan menjadi DKC. Setiap DKC direpresentasikan melalui *topic name, topic summary, topic keywords, topic representation*, dan dokumen pendukung. Evaluasi pada tahap ini tidak hanya memeriksa keberadaan atribut tersebut, tetapi juga mengukur hubungan semantik antara *topic representation* dengan *keywords, summary*, dan dokumen sumber.

Dari 173 *domain knowledge* yang dihasilkan, rata-rata *Keyword–Representation Similarity* adalah 0,8192, sedangkan *Representation–Summary Similarity* sebesar 0,7921 dan *Representation–Document Similarity* sebesar 0,6626. *Median* masing-masing ukuran adalah 0,8215, 0,8033, dan 0,6723.

Tabel 5. Evaluasi *Semantic Similarity* pada *Domain Knowledge Collection*

Hubungan Semantik	Mean	Median	Std. Dev.
Keyword–Representation	0,8192	0,8215	0,0620
Representation–Summary	0,7921	0,8033	0,0685
Representation–Document	0,6626	0,6723	0,0845

Nilai tersebut menunjukkan bahwa *topic representation* memiliki kedekatan semantik yang tinggi dengan konsep utama (*keywords* dan *summary*), tetapi memiliki kedekatan yang lebih rendah terhadap dokumen individual. Pola tersebut sesuai dengan fungsi *representation* sebagai abstraksi dari kumpulan dokumen, karena *representation* mempertahankan konsep utama tanpa harus mereproduksi keseluruhan karakteristik setiap dokumen sumber.

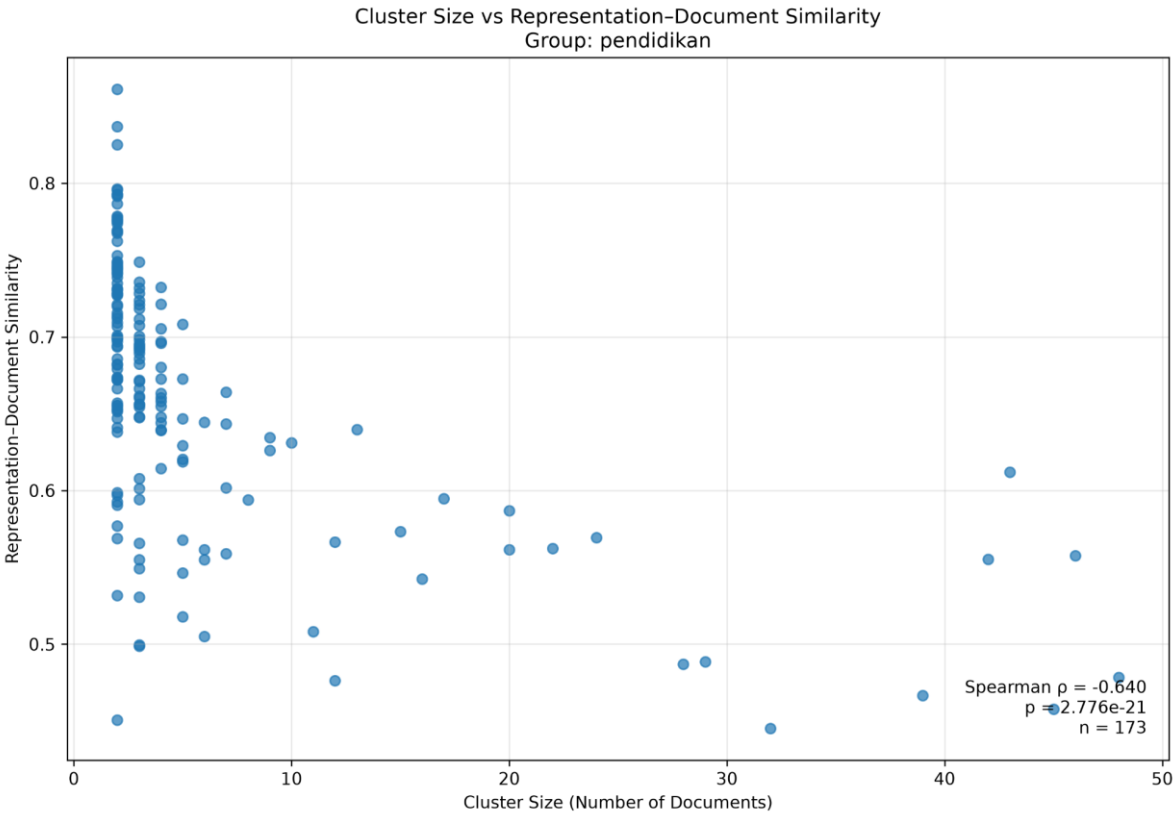
Tabel 6. Perbandingan Evaluasi *Semantic Similarity* Antar Ukuran *Cluster*

Ukuran *cluster*	Jumlah *cluster*	*Keyword–Rep.*	*Rep.–Summary*	*Rep.–Document*
Kecil (≤9)	152	0,8183	0,7922	0,6794
Menengah (10–19)	8	0,8228	0,7693	0,5666
Besar (≥20)	13	0,8280	0,8048	0,5253

Hubungan tersebut menjadi lebih jelas ketika dianalisis berdasarkan ukuran *cluster*. Pada 13 *cluster* besar, rata-rata *Keyword–Representation Similarity* mencapai sekitar 0,828, sedangkan *Representation–Summary Similarity* sekitar 0,805. Sebaliknya, *Representation–Document Similarity* turun menjadi sekitar 0,525. Pada cluster kecil, nilai *Representation–Document Similarity* relatif lebih tinggi, sedangkan pada *cluster* menengah berada di antara keduanya.

Temuan tersebut diperkuat oleh analisis korelasi antara ukuran *cluster* dan ketiga ukuran *similarity*. Tidak ditemukan hubungan yang signifikan antara ukuran *cluster* dan *Keyword–Representation Similarity* (Spearman $\rho = 0{,}0109$; $p = 0{,}8869$). Hal serupa terjadi pada *Representation–Summary Similarity* ($\rho = -0{,}1084$; $p = 0{,}1559$). Sebaliknya, terdapat hubungan negatif yang kuat antara ukuran *cluster* dan *Representation–Document Similarity* ($\rho = -0{,}6395$; $p < 0{,}001$).

Submitted : 13-06-2026| Reviewed : 20-06-2026 | Accepted : 30-06-2026



Gambar 2. Hubungan Ukuran *Cluster* dan *Representation–Document Similarity* pada Domain Pendidikan

Peningkatan ukuran *cluster* tidak secara signifikan mengubah kedekatan representation terhadap *keywords* maupun *summary*. Namun, semakin besar *cluster*, semakin rendah kedekatan representation terhadap dokumen individual. Gambar 2. memperlihatkan pola tersebut secara visual, dengan korelasi Spearman −0,640 pada 173 *cluster*. *Cluster* yang besar cenderung memiliki *coherence* lebih rendah dan *Representation–Document Similarity* yang lebih rendah. Dengan demikian, abstraksi yang dihasilkan dari *cluster* besar dapat menjadi lebih umum karena mencakup wilayah semantik yang lebih luas. Kondisi ini menunjukkan bahwa ukuran dan homogenitas *cluster* tetap menjadi faktor penting dalam menentukan kualitas DKC.

3.4 Pembahasan

Hasil eksperimen menunjukkan bahwa PTF mampu membentuk rangkaian proses dari *semantic retrieval* menuju *progressive document clustering* dan selanjutnya menghasilkan DKC. Evaluasi tidak menunjukkan bahwa seluruh *cluster* memiliki kualitas yang sama. Sebaliknya, terdapat pola yang konsisten antara ukuran *cluster*, koherensi semantik, dan karakteristik representation.

Pada tahap *clustering*, nilai *mean coherence* sebesar 0,8573 menunjukkan bahwa secara umum dokumen dalam *cluster* memiliki kedekatan semantik yang tinggi. Namun, korelasi Spearman sebesar −0,9269 menunjukkan bahwa nilai tersebut sangat dipengaruhi oleh ukuran *cluster*. *Cluster* kecil cenderung memiliki homogenitas tinggi, sedangkan *cluster* besar memiliki cakupan semantik yang lebih luas. Temuan *inter-cluster similarity* yang tinggi pada beberapa pasangan *cluster* besar juga menunjukkan bahwa sebagian wilayah konsep masih memiliki *overlap*.

Temuan tersebut justru memperjelas fungsi PTF sebagai proses *progressive topic formation*. *Cluster* tidak langsung dianggap sebagai pengetahuan final, tetapi menjadi input untuk membangun representasi pengetahuan. Pada tahap DKC, *topic representation* memiliki *similarity* yang tinggi terhadap *keywords* dan *summary*, sementara *similarity* terhadap dokumen individual lebih rendah. Korelasi negatif antara ukuran *cluster* dan *Representation–Document Similarity* menunjukkan bahwa *representation* semakin berperan sebagai abstraksi ketika jumlah dokumen sumber meningkat.

Karakteristik tersebut membedakan tujuan PTF dari *topic modeling* konvensional. LDA membangun topik berdasarkan distribusi probabilistik kata terhadap dokumen dan topik, sedangkan BERTopic memanfaatkan

representasi *embedding* dokumen untuk membentuk dan merepresentasikan topik. PTF menggunakan *clustering* sebagai salah satu tahap pembentukan topik, tetapi melanjutkannya dengan proses konstruksi *knowledge object* yang menggabungkan *summary*, *keywords*, *representation*, *embedding*, dan dokumen pendukung. Dengan demikian, perbedaan utama bukan semata-mata pada teknik *clustering*, melainkan pada unit keluaran dan tujuan proses.

Hasil eksperimen juga menunjukkan bahwa penggunaan LLM pada PTF tidak seharusnya dipahami sebagai mekanisme yang secara otomatis memperbaiki kualitas *cluster*. Kualitas representation tetap dipengaruhi oleh kualitas dokumen yang masuk ke *cluster*. *Cluster* dengan *coherence* rendah dapat menghasilkan representation yang tetap konsisten dengan *keywords* dan *summary*, tetapi memiliki kedekatan yang lebih rendah terhadap dokumen sumber. Kondisi ini menunjukkan adanya *trade-off* antara cakupan informasi dan spesifisitas abstraksi.

Dengan demikian, kontribusi PTF terletak pada integrasi proses *retrieval, clustering*, dan abstraksi pengetahuan dalam satu *framework. Semantic Domain Prototype* membatasi ruang domain, *clustering* membentuk kelompok dokumen, sedangkan *Domain Knowledge Construction* mengubah kelompok tersebut menjadi *knowledge object* yang dapat digunakan kembali. Hasil eksperimen mendukung kemampuan framework dalam menghasilkan representasi pengetahuan yang terstruktur, tetapi sekaligus menunjukkan bahwa kualitas DKC perlu dipertimbangkan bersama *homogenitas cluster* dan ukuran kelompok dokumen.

4. Kesimpulan

Penelitian ini mengusulkan PTF sebagai *framework* untuk membangun DKC dari koleksi dokumen tidak terstruktur. PTF mengintegrasikan *Semantic Domain Prototype*, *semantic retrieval, progressive document clustering,* dan *Domain Knowledge Construction* untuk membentuk *knowledge object* yang memadukan nama topik, ringkasan topik, kata kunci, representasi topik, dan dokumen pendukung.

Eksperimen menggunakan 599.059 artikel Wikipedia Bahasa Indonesia sebagai koleksi dokumen dan 1.000 dokumen hasil *semantic retrieval* sebagai masukan proses *clustering* pada Domain Pendidikan. Proses tersebut menghasilkan 173 *cluster* tanpa *noise*, dengan *mean semantic coherence* sebesar 0,8573. Analisis menunjukkan hubungan negatif yang sangat kuat antara ukuran *cluster* dan *coherence* (Spearman $\rho = -0{,}9269$; $p < 0{,}001$), yang menunjukkan bahwa *cluster* berukuran kecil cenderung lebih homogen, sedangkan *cluster* berukuran besar mencakup ruang semantik yang lebih luas.

Pada tahap pembentukan DKC, *topic representation* memiliki kedekatan semantik yang tinggi dengan *keywords* dan *topic summary*, masing-masing dengan rata-rata *similarity* sebesar 0,8192 dan 0,7921. Sementara itu, *Representation–Document Similarity* memiliki rata-rata 0,6626 dan menunjukkan korelasi negatif yang kuat terhadap ukuran *cluster* (Spearman $\rho = -0{,}6395$; $p < 0{,}001$). Pola ini menunjukkan bahwa *topic representation* berfungsi sebagai abstraksi konseptual yang mempertahankan informasi utama suatu *cluster* tanpa sekadar merepresentasikan dokumen individual.

Hasil tersebut menunjukkan bahwa PTF berhasil mengembangkan proses pembentukan topik menjadi konstruksi pengetahuan domain secara bertahap, dengan DKC sebagai keluaran terstruktur yang dapat digunakan kembali. Dengan demikian, kontribusi utama penelitian ini bukan hanya pada pembentukan *cluster* dokumen, tetapi pada mekanisme untuk mentransformasikan *cluster* tersebut menjadi *knowledge object* yang memiliki konteks semantik dan referensi dokumen pendukung.

Temuan mengenai hubungan ukuran *cluster* dengan *semantic coherence* dan *Representation–Document Similarity* juga membuka peluang untuk mengembangkan mekanisme *adaptive domain refinement*, khususnya untuk mengidentifikasi dan memproses kembali *cluster* berukuran besar yang memiliki koherensi rendah atau *overlap* semantik tinggi. Selain itu, DKC yang dihasilkan dapat dikembangkan lebih lanjut melalui integrasi dengan *knowledge graph* dan *Retrieval-Augmented Generation* untuk menguji pemanfaatannya sebagai lapisan pengetahuan pada sistem berbasis AI.

Daftar Rujukan

[1] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 9459–9474.

[2] Y. Gao *et al.*, “Retrieval-Augmented Generation for Large Language Models: A Survey,” Mar. 27, 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 11, 2022, *arXiv*: arXiv:2203.05794. doi: 10.48550/arXiv.2203.05794.

[5] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[6] W. Hardi, "PENGELOMPOKAN TOPIK DOKUMEN BERBASIS TEXT MINING DENGAN ALGORITME K-MEANS: STUDI KASUS PADA DOKUMEN KEDUTAAN BESAR AUSTRALIA JAKARTA," vol. 21, no. 1, 2019.

[7] G. Patidar, A. Singh, and D. Singh, "An Approach for Document Clustering using Agglomerative Clustering and Hebbian-type Neural Network," *IJCA*, vol. 75, no. 9, pp. 17–22, Aug. 2013, doi: 10.5120/13139-0532.

[8] A. Eklund, M. Forsman, and F. Drewes, "An Empirical Configuration Study of a Common Document Clustering Pipeline," *Northern European Journal of Language Technology*, vol. 9, 2023, doi: 10.3384/nejlt.2000-1533.2023.4396.

[9] R. Gusmansyah, R. Rahmaddeni, R. Rohid, A. Rivaldi, and S. Daulay, "Perbandingan Metode Machine Learning untuk Klastering Penerima Bantuan Pendidikan Siswa," *Jurnal Pustaka AI*, vol. 5, no. 2, pp. 193–203, Aug. 2025, doi: 10.55382/jurnalpustakaai.v5i2.1139.

[10] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds., Berlin, Heidelberg: Springer, 2013, pp. 160–172. doi: 10.1007/978-3-642-37456-2_14.

[11] R. J. G. B. Campello, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *WIREs Data Min & Knowl*, vol. 10, no. 2, p. e1343, Mar. 2020, doi: 10.1002/widm.1343.

[12] "indonesian-nlp/wikipedia-id · Datasets at Hugging Face." Accessed: Apr. 07, 2026. [Online]. Available: https://huggingface.co/datasets/indonesian-nlp/wikipedia-id

[13] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing Semantic Coherence in Topic Models".

[14] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," *Heliyon*, vol. 11, no. 2, p. e41953, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41953.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

Submitted : 13-06-2026| Reviewed : 20-06-2026 | Accepted : 30-06-2026