Volume 8 Nomor 1 Januari 2026 JJEEE e-ISSN : 2715-0887
p-ISSN : 2654-7813

Jambura Journal of Electrical and Electronics Engineering

Optimasi *Embedding* IndoBERT dengan *Ditto Whitening* untuk Pengukuran Kesamaan Judul Penelitian

Optimizing of IndoBERT Embedding with Ditto Whitening for Measuring Research Title Similarity

Rezqiwati Ishak
Progran Studi Teknik Informatika
Universitas Ichsan Gorontalo
Gorontalo, Indonesia
rezqi.uig@gmail.com

Amiruddin*
Program Studi Teknik Informatika
Universitas Ichsan Goorntalo
Gorontalo, Indonesia
amier.76@gmail.com

Diterima : November 2025
Disetujui : Januari 2026
Dipublikasi : Januari 2026

Abstrak*—Pengukuran kesamaan judul penelitian merupakan elemen penting dalam menjaga orisinalitas karya ilmiah serta mencegah duplikasi topik di lingkungan akademik. Namun, *embedding* IndoBERT sebagai model bahasa Indonesia diketahui mengalami masalah anisotropi, yang menyebabkan sebagian besar judul tampak memiliki nilai kemiripan tinggi meskipun berbeda secara semantik. Penelitian ini bertujuan untuk mengoptimalkan kualitas *embedding* IndoBERT melalui *Ditto Whitening*. Selain itu, penelitian ini juga mengevaluasi dampaknya terhadap pengukuran kesamaan judul penelitian. Dataset yang digunakan terdiri dari 7.785 judul tugas akhir mahasiswa dari enam rumpun ilmu, yang diproses menggunakan teknik *mean pooling* dan normalisasi L2 sebelum dan sesudah *whitening*. Evaluasi dilakukan secara intrinsik dengan mengukur tingkat isotropi, distribusi *cosine similarity*, bias global terhadap vektor *mean*, dan fenomena *hubness*, serta didukung oleh visualisasi ruang *embedding* menggunakan t-SNE, UMAP, dan *heatmap cosine similarity*. Hasil eksperimen menunjukkan peningkatan signifikan pada kualitas *embedding*, ditunjukkan oleh penurunan *Cosine Pair Mean* dari 0.559 menjadi −0.000145, penurunan *MeanCos-to-Mean* dari 0.748 menjadi 0.0068, serta penurunan *Hubness Skew* dari 1.60 menjadi 0.68. Nilai isotropi *embedding* juga meningkat secara substansial, menunjukkan distribusi vektor yang lebih merata. Temuan ini membuktikan bahwa *Ditto Whitening* efektif dalam meningkatkan isotropi *embedding* IndoBERT dan secara langsung berdampak pada peningkatan akurasi sistem deteksi kemiripan judul serta pencarian dokumen akademik, sehingga relevan untuk mendukung manajemen topik dan penjaminan mutu penelitian di perguruan tinggi.

Kata Kunci—IndoBERT; Whitening; Embedding; Kesamaan Semantik; Judul Penelitian

Abstract—Measuring the semantic similarity of research titles is a crucial component in maintaining academic originality and preventing topic duplication in higher education. However, IndoBERT embeddings, as a pretrained Indonesian language model, are known to suffer from anisotropy, causing many titles to exhibit high similarity scores despite being semantically distinct. This study aims to optimize the quality of IndoBERT embeddings through Ditto Whitening and to evaluate its impact on research title similarity measurement. The dataset comprises 7.785 undergraduate thesis titles collected from six disciplinary domains and processed using mean pooling and L2 normalization before and after whitening. An intrinsic evaluation was conducted by assessing embedding isotropy, cosine similarity distribution, global bias toward the mean vector, and hubness phenomena, supported by embedding space visualizations using t-SNE, UMAP, and cosine similarity heatmaps. Experimental results demonstrate substantial improvements in embedding quality, indicated by a reduction in Cosine Pair Mean from 0.559 to −0.000145, a decrease in MeanCos-to-Mean from 0.748 to 0.0068, and a reduction in Hubness Skew from 1.60 to 0.68. The isotropy of the embeddings also increased markedly, reflecting a more uniform vector distribution. These findings confirm that Ditto Whitening effectively improves the isotropy of IndoBERT embeddings and directly enhances the accuracy of research title similarity detection and academic document retrieval systems, thereby supporting topic management and research quality assurance in higher education.

Keywords—IndoBERT; Whitening; Embedding; Semantic Similarity; Research Titles

I. PENDAHULUAN

Pengukuran kesamaan semantik pada judul penelitian merupakan kebutuhan penting dalam proses penjaminan mutu akademik, khususnya untuk mencegah duplikasi topik, mendeteksi kemiripan tematik, serta membantu menilai orisinalitas penelitian mahasiswa. Judul penelitian termasuk kategori *short scientific text*, yaitu teks ilmiah pendek yang memiliki struktur linguistik ringkas, padat makna, dan sering kali menggunakan istilah teknis yang serupa lintas rumpun ilmu. Kondisi ini membuat sistem berbasis pencocokan kata atau pendekatan statistik tradisional kurang efektif, karena tidak mampu menangkap representasi semantik yang mendalam[1]. Oleh karena itu, penggunaan *pre-trained language models* seperti BERT dan turunannya menjadi pilihan utama dalam representasi semantik teks pendek[2][3].

Representasi semantik teks pada pemrosesan bahasa alami umumnya dibangun dalam bentuk *embedding*, yaitu vektor numerik berdimensi tinggi yang merepresentasikan makna suatu teks dalam ruang vektor. *Embedding* memungkinkan pengukuran kemiripan semantik antar teks menggunakan metrik seperti *cosine similarity*. Meskipun *embedding* berbasis BERT mampu menangkap konteks secara mendalam, berbagai studi menunjukkan bahwa *embedding* tersebut cenderung mengalami fenomena anisotropi, yaitu distribusi vektor yang tidak merata dalam ruang *embedding* sehingga sebagian besar pasangan teks memiliki nilai *cosine similarity* yang relatif tinggi meskipun secara semantik tidak berkaitan[4]. Kondisi ini mengurangi kemampuan model dalam membedakan konteks, terutama pada domain teks pendek seperti judul penelitian. Penelitian lain juga menunjukkan bahwa anisotropi memicu fenomena *hubness*, yaitu kecenderungan sebagian vektor menjadi tetangga dekat bagi banyak vektor lain secara tidak proporsional, sehingga menurunkan kualitas pemetaan semantik secara keseluruhan[5].

Permasalahan tersebut menjadi semakin kompleks pada konteks judul tugas akhir dan skripsi mahasiswa di Indonesia. Judul-judul ini umumnya disusun dengan pola formal yang relatif seragam dan memuat istilah teknis umum seperti "analisis", "pengaruh", atau "implementasi" yang digunakan secara luas lintas rumpun ilmu, termasuk ekonomi, hukum, teknik, dan ilmu komputer. Karakteristik ini memperbesar kecenderungan *embedding* untuk terkonsentrasi pada arah vektor tertentu, sehingga memperparah efek anisotropi dan bias kemiripan global pada ruang *embedding*. Untuk mengatasi permasalahan anisotropi, sejumlah pendekatan *whitening* dan normalisasi *embedding* telah dikembangkan, seperti PCA *Whitening*[6], *all-but-the-top*[7], *BERT-Flow*[5] dan *PromptBERT*[8]. Pendekatan-pendekatan tersebut terbukti mampu meningkatkan isotropi *embedding* dan memperbaiki performa pengukuran kemiripan pada beberapa *benchmark* internasional, seperti STS-B dan SICK-R. Namun, sebagian besar penelitian tersebut dilakukan pada bahasa Inggris, sementara kajian mengenai peningkatan kualitas *embedding* pada bahasa Indonesia masih terbatas[9]. Selain itu, penerapan metode *whitening* pada model IndoBERT, khususnya dalam konteks teks akademik berbahasa Indonesia, belum banyak dieksplorasi secara sistematis.

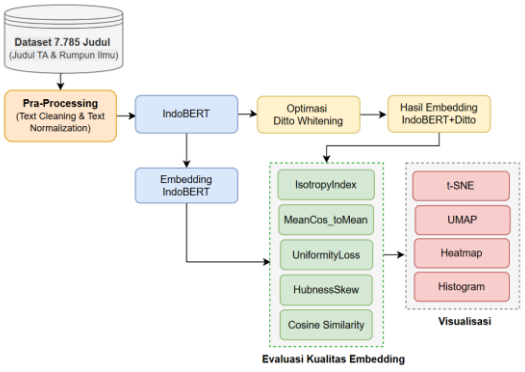
Salah satu pendekatan yang relatif baru dan menunjukkan performa menjanjikan adalah *Ditto Whitening*, yaitu metode optimasi *embedding* berbasis transformasi *linear* yang dirancang untuk menghasilkan distribusi vektor yang lebih isotropik tanpa mengubah struktur semantik internal *embedding*[10]. Penelitian awal menunjukkan bahwa *Ditto Whitening* mampu menurunkan *cosine-pair* bias dan meningkatkan keseragaman distribusi *embedding* pada berbagai skenario representasi teks pendek. Namun, hingga saat ini, belum terdapat kajian yang secara spesifik meneliti efektivitas *Ditto Whitening* dalam meningkatkan kualitas *embedding* IndoBERT, khususnya dalam konteks pengukuran kesamaan judul penelitian berbahasa Indonesia.

Berdasarkan kondisi tersebut, penelitian ini berfokus pada optimalisasi *embedding* IndoBERT *menggunakan Ditto Whitening* untuk meningkatkan akurasi pengukuran kesamaan semantik judul penelitian. Evaluasi dilakukan melalui serangkaian metrik kualitas *embedding* secara intrinsik, meliputi *isotropy index, uniformity loss, hubness skew*, dan distribusi *cosine similarity*, serta didukung oleh visualisasi ruang *embedding* menggunakan t-SNE, UMAP, dan *heatmap cosine similarity*. Penelitian ini diharapkan memberikan kontribusi berupa: (1) analisis diagnostik karakteristik anisotropi *embedding* IndoBERT pada judul penelitian berbahasa Indonesia, (2) pembuktian empiris efektivitas *Ditto Whitening* sebagai metode optimasi *embedding* pada model bahasa Indonesia, dan (3) penyediaan dasar metodologis bagi pengembangan sistem deteksi kemiripan judul penelitian yang lebih akurat untuk kebutuhan akademik.

II. METODE

A. Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan utama, yaitu pengumpulan dataset judul penelitian, pra-pemrosesan teks, pembentukan *embedding* menggunakan model IndoBERT, optimasi *embedding* menggunakan *Ditto Whitening*, serta evaluasi kualitas *embedding* secara intrinsik. Alur penelitian dirancang untuk mengevaluasi pengaruh *Ditto Whitening* terhadap karakteristik geometri ruang *embedding* secara sistematis dan terukur. Tahapan penelitian diuraikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

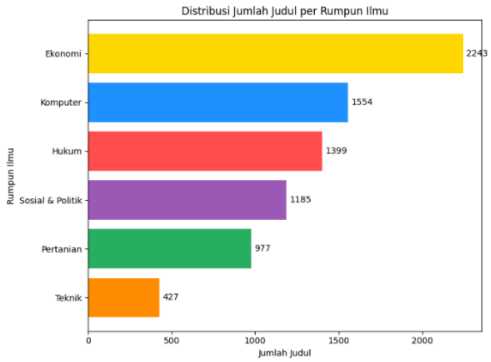
B. Dataset dan Pra-pemrosesan

Dataset yang digunakan dalam penelitian ini terdiri dari 7.785 judul tugas akhir mahasiswa yang diperoleh dari Sistem Informasi Akademik Universitas Ichsan Gorontalo dan mencakup enam rumpun ilmu, yaitu Ekonomi, Komputer, Hukum, Sosial dan Politik, Pertanian, serta Teknik. Dataset ini dipilih karena merepresentasikan konteks teks akademik pendek yang umum digunakan dalam penilaian orisinalitas penelitian mahasiswa[11]. Tahap pra-pemrosesan dilakukan untuk menyiapkan teks agar sesuai dengan kebutuhan model berbasis transformer. Proses pra-pemrosesan meliputi konversi huruf menjadi huruf kecil (*case folding*), pembersihan karakter *non-alfanumeric*, normalisasi spasi, serta penghapusan judul dengan jumlah kata kurang dari sepuluh. Penghapusan judul pendek (<10 kata) dilakukan karena judul yang terlalu singkat cenderung mengandung informasi semantik yang terbatas dan berpotensi menghasilkan *embedding* yang tidak stabil. Keputusan ini diambil untuk menjaga konsistensi kualitas representasi semantik, meskipun berimplikasi pada pengurangan sebagian kecil data, yang menjadi salah satu batasan penelitian ini. Tahapan ini mengacu pada praktik prapemrosesan standar dalam pemodelan transformer [12][13]. Sampel dataset setelah dilakukan pra-pemrosesan dapat dilihat pada Tabel 1.

TABEL 1. SAMPEL DATASET PENELITIAN

No	Judul TA	Rumpun Ilmu
1	analisis nilai tambah pada agroindustri tepung...	Pertanian
2	studi pembuatan donat dengan penambahan tepung...	Pertanian
3	penerapan penjatuhan sanksi tindak pidana perjudian di pengadilan negeri gororntalo	Hukum
4	efektivitas pemungutan pajak bumi dan bangunan...	Sosial & Politik
5	pembinaan narapidana penyalahgunaan narkotika ...	Hukum
6	pengaruh persepsi harga terhadap keputusan pem...	Ekonomi
7	aplikasi untuk diagnosa penyakit menular seksual pada manusia dengan metode fuzzy logic	Komputer
8	peran lembaga pemberdayaan masyarakat dalam pa...	Sosial & Politik
9	analisis tingkat resiko kredit pada bank bumn ...	Ekonomi
10	pengaruh perilaku konsumen terhadap keputusan ...	Ekonomi
...	...	...
7.785	perencanaan kantor dinas tata kota dan pertamanan kota gorontalo dengan pendekatan arsitektur hijau	Teknik

Dataset yang digunakan mencakup enam rumpun ilmu, yaitu Ekonomi, Komputer, Hukum, Sosial & Politik, Pertanian, dan Teknik. Distribusi per rumpun ilmu data dapat dilihat pada Gambar 2.



Gambar 2. Distribusi Dataset berdasarkan Rumpun Ilmu

C. Model IndoBERT dan Lingkungan Eksperimen

Embedding Model bahasa yang digunakan dalam penelitian ini adalah IndoBERT-base-p2 dengan nama repositori indobenchmark/indobert-base-p2, yang tersedia secara publik melalui platform *HuggingFace*. Model ini dipilih karena telah dilatih pada korpus besar berbahasa Indonesia dan banyak digunakan sebagai *baseline* dalam penelitian pemrosesan bahasa alami berbahasa Indonesia.

Eksperimen dilakukan menggunakan *framework PyTorch* dan *library Transformers* dari *HuggingFace*. Seluruh proses komputasi dijalankan pada perangkat dengan akselerasi GPU. Parameter dan konfigurasi utama eksperimen disajikan pada Tabel 2.

TABEL 2. KONFIGURASI PARAMETER EKSPERIMEN

No	Komponen	Konfigurasi
1	*Model*	indobenchmark/indobert-base-p2
2	*Framework*	PyTorch
3	*Library*	*Transformers* (*HuggingFace*)
4	*Batch size*	32
5	Panjang token maksimum	128
6	*Pooling*	*Mean pooling*
7	Normalisasi	*L2 normalization*
8	Perangkat	GPU

D. Pembentukan Embedding IndoBERT

Embedding judul penelitian diperoleh dari *last hidden state* model IndoBERT dengan menggunakan teknik *mean pooling*, yaitu dengan merata-ratakan vektor token berdasarkan *attention mask*. *Mean pooling* dipilih karena mampu menghasilkan representasi kalimat yang lebih stabil dan tidak terlalu bergantung pada satu token tertentu, dibandingkan penggunaan token [CLS] yang diketahui sensitif terhadap anisotropi pada *embedding* berbasis BERT. *Embedding* yang dihasilkan kemudian dinormalisasi menggunakan *L2 normalization* untuk menyetarakan skala vektor sehingga pengukuran *cosine similarity* menjadi lebih *reliabel*. Hasil tahap ini menghasilkan *embedding baseline* dengan konfigurasi IndoBERT + L2, yang digunakan sebagai pembanding utama dalam evaluasi[3]. *Mean pooling* dapat dihitung menggunakan persamaan (1).

$$e = \frac{\sum_{i=1}^{n}(h_i \cdot m_i)}{\sum_{i=1}^{n} m_i} \quad (1)$$

Di mana :

h_i = *hidden state token* ke-i,
m_i = *attention mask*,
n = jumlah token,
e = *embedding* kalimat.

Selanjutnya dilakukan normalisasi L2 untuk menyetarakan skala antar *embedding* agar pengukuran *cosine similarity* lebih reliabel[14]. Normalisasi ini dapat dihitung menggunakan persamaan (2).

$$e_{norm} = \frac{e}{|e|} \quad (2)$$

Di mana :

enorm = *embedding* ter-normalisasi
‖e‖ = norma L2

Hasil *embedding* IndoBERT ini menghasilkan matriks vektor berdimensi tinggi yaitu 7.785 x 768 dengan sampel ditujukkan pada Tabel 3.

TABEL 3. SAMPEL *EMBEDDING* IndoBERT

No	Dim_0	Dim_1	Dim_2	...	dim_767
1	0.003037	0.041321	-0.0152	...	-0.03805
2	0.020828	0.028831	-0.01141	...	-0.02109
3	-0.00527	0.015855	0.014384	...	-0.05944
4	0.006347	0.011476	0.0145	...	-0.06057
5	-0.04699	0.027351	0.013718	...	-0.05123
6	0.0306	0.032365	-0.02992	...	-0.0297
7	0.017609	0.025515	0.003072	...	-0.01917

No	Dim_0	Dim_1	Dim_2	...	dim_767
8	-0.00877	0.009185	0.025582	...	-0.0588
9	0.006021	0.011959	0.044129	...	-0.05564
...	...	...	**...**	...	...
7.785	0.011415	0.019605	0.001535	...	0.030413

E. Optimasi Embedding Menggunakan Ditto Whitening

Optimasi *embedding* dilakukan menggunakan *Ditto Whitening*, yaitu metode transformasi *linear* yang bertujuan meningkatkan isotropi *embedding* dengan cara menghilangkan korelasi antar dimensi tanpa mengubah struktur semantik internal *embedding*. Proses *Ditto Whitening* mencakup beberapa langkah utama, yaitu *mean-centering*, pembentukan matriks kovarians, dekomposisi *eigen*, transformasi *whitening*, dan normalisasi akhir. Proses ini menghasilkan *embedding* IndoBERT + *Ditto Whitening* yang kemudian dievaluasi secara intrinsik[10]. Setiap tahap diperlukan agar *embedding* memiliki variansi antar dimensi yang lebih merata, sehingga hasil pengukuran kesamaan semantik lebih akurat[10].

1). *Mean-Centering*

Tahap ini bertujuan menghilangkan pengaruh nilai rata-rata terhadap distribusi *embedding*, sehingga struktur kovarian dapat dihitung secara tepat[15]. *Mean-centering* dapat dihitung menggunakan persamaan (3).

$$X_c = X - \mu \tag{3}$$

Di mana:

- X_c = Matriks embedding yang telah dipusatkan (*zero-centered matrix*), di mana titik pusat data digeser ke nol.
- X = Matriks input embedding asli (dari IndoBERT) sebelum diproses.
- μ = Vektor rata-rata global (*global mean vector*) dari seluruh sampel judul dalam dataset.

2) *Covariance Matrix*

Tahap ini menghitung hubungan antar dimensi *embedding*. Matriks kovarian digunakan sebagai dasar untuk menentukan struktur *whitening*[16]. Nilai kovarian dapat dihitung menggunakan persamaan (4)[17].

$$C = \frac{1}{N-1} X_c^\top X_c \tag{4}$$

Di mana:

- N = Jumlah total sampel (judul penelitian) dalam dataset.
- X_c = Matriks *embedding* (IndoBERT) yang telah dipusatkan (*zero-centered*), di mana rata-rata setiap fitur telah dikurangi menjadi nol.
- T = Simbol operasi *transpose* matriks (menukar baris menjadi kolom).

3) *Eigen Decomposition*

Tahap ini mengekstraksi *eigenvalue* dan *eigenvector* dari matriks kovarian sebagai dasar transformasi *whitening*[17]. Proses ini dapat dihitung menggunakan persamaan (5).

$$C = U \Lambda U^\top \tag{5}$$

Di mana:

- C = Matriks kovarians sampel yang akan didekomposisi.
- U = Matriks ortogonal yang berisi vektor *eigen* (*eigenvectors*) dari matriks kovarians.
- Λ = Matriks diagonal yang berisi nilai *eigen* (*eigenvalues*), yang merepresentasikan besaran variansi di setiap arah vektor *eigen*.
- T = Simbol operasi *transpose*.

4) *Whitening Transform*

Transformasi *whitening* bertujuan menghilangkan korelasi antar dimensi embedding sehingga distribusi menjadi lebih isotropik[16][6]. *Ditto Whitening* dapat dihitung menggunakan persamaan (6) dan (7).

$$W = U \Lambda^{-\frac{1}{2}} U^\top \tag{6}$$

$$Z = X_c W \tag{7}$$

Di mana:

- X_c = Matriks *embedding* IndoBERT yang telah dipusatkan (*zero-centered*).
- W = Matriks transformasi *Ditto Whitening*.
- U = Matriks vektor *eigen* yang diperoleh dari dekomposisi nilai *eigen*.
- Λ = Matriks diagonal yang berisi nilai *eigen*.
- Z = Representasi *embedding* akhir setelah proses optimasi (*whitened embeddings*).

5) Normalisasi Akhir

Embedding hasil *whitening* kemudian dinormalisasi kembali agar stabil dalam perhitungan *cosine similarity*[18]. Normalisasi akhir dapat dihitung menggunakan persamaan (8).

$$Cz_{norm} = \frac{Z}{|Z|} \tag{8}$$

Di mana:

- Z = Representasi embedding hasil transformasi *Whitening* (sebelum dinormalisasi).
- |Z| = Magnitudo (panjang) atau norma L2 (*L2-norm*) dari vektor Z.

F. Baseline Pembanding

Untuk menilai efektivitas *Ditto Whitening* secara objektif, penelitian ini menggunakan dua konfigurasi *baseline* utama, yaitu:

1) IndoBERT + *L2 normalization*, sebagai baseline dasar tanpa optimasi distribusi embedding.
2) IndoBERT + *Ditto Whitening*, sebagai model yang diusulkan dalam penelitian ini.

Perbandingan dilakukan secara langsung pada kedua konfigurasi tersebut menggunakan metrik evaluasi intrinsik yang sama, sehingga dampak *Ditto Whitening* terhadap kualitas *embedding* dapat diamati secara jelas. Konfigurasi *baseline* ini dipilih untuk menjaga fokus penelitian pada analisis geometri *embedding* IndoBERT.

G. Evaluasi Kualitas Embedding

Evaluasi kualitas *embedding* dilakukan menggunakan sejumlah metrik intrinsik untuk mengukur tingkat isotropi, bias global, dan distribusi ruang *embedding*. Metrik-metrik ini umum digunakan dalam penelitian *embedding* modern [19][20].

1) *Isotropy Index*

Metrik ini mengukur keseimbangan variansi antar dimensi *embedding*. Semakin mendekati 1, semakin isotropik ruang *embedding*[7]. *Isotropy Index* dapat dihitung menggunakan persamaan (9).

$$C\text{IsoIndex} = \frac{\lambda_{min}}{\lambda_{max}} \tag{9}$$

Di mana:

$I_{IsoIndex}$ = indeks isotropi ruang *embedding* ($0 < (I_{iso})$;
λ_{min} = nilai *eigen* terkecil dari matriks kovarians;
λ_{max} = nilai *eigen* terbesar dari matriks kovarians;

2) *MeanCos-to-Mean*

Metrik ini mengukur kedekatan *embedding* terhadap vektor mean. Nilai yang tinggi mengindikasikan bias *cosine* yang kuat[4]. *MeanCos-to-Mean* dapat dihitung menggunakan Persamaan (10).

$$MeanCosMean = \frac{1}{N}\sum_{i=1}^{N} \cos(\boldsymbol{x_i}, \boldsymbol{\mu}) \tag{10}$$

Di mana:

N = Jumlah total data dalam pengujian.
x_i = Vektor embedding untuk judul ke-i.
μ = Vektor rata-rata global dari seluruh data.

3) *Uniformity Loss*

Metrik ini mengukur tingkat sebaran *embedding* dalam ruang *manifold*, semakin rendah nilainya, semakin seragam penyebaran *embedding*[19]. *Uniformity Loss* dapat dihitung menggunakan Persamaan (11).

$$\mathcal{L}_{uni} = \log E_{i,j}\left[e^{-2|x_i - x_j|^2}\right] \tag{11}$$

Di mana:

$E_{i,j}$ = Nilai harapan (*Expectation*) terhadap pasangan indeks acak i dan j.
x_i, x_j = Vektor *embedding* dari sampel ke-i dan ke-j yang telah dinormalisasi (berada pada *hypersphere*).
$\| . \|$ = Jarak Euclidean antar dua vektor *embedding*.
-2 = Parameter suhu (*temperature*) standar yang digunakan dalam fungsi kernel Gaussian.

4) *Hubness Skew*

Skewness dari distribusi jumlah *nearest neighbors* mengukur apakah ada *hub vectors*. Nilai rendah menunjukkan ruang *embedding* lebih merata[4]. *Hubness Skew* dapat dihitung menggunakan Persamaan (12).

$$\text{HubnessSkew} = \text{Skew}(c_1, c_2, \ldots, c_N) \tag{12}$$

Di mana:

Skew = Fungsi statistik untuk mengukur kemencengan distribusi (*skewness*).
c_i = Nilai *k-occurrence* (frekuensi sampel ke-i muncul sebagai tetangga terdekat bagi sampel lain).
$c_1, \ldots, c_n$ = Himpunan nilai *k-occurrence* untuk seluruh data.

5) *Cosine Pair Mean*

Metrik ini mengukur kecenderungan kemiripan global antar *embedding*. Nilai tinggi mengindikasikan bias *cosine*[4]. *Cosine Pair Mean* dapat dihitung menggunakan Persamaan (13).

$$CosPairMean = \frac{1}{M}\sum_{(i,j)} \cos(x_i, x_j) \tag{13}$$

Di mana:

M = Jumlah pasangan judul yang dipilih secara acak untuk evaluasi.
x_i, x_j = Vektor *embedding* dari pasangan judul ke-i dan ke-j yang dibandingkan.

H. Visualisasi Ruang Embedding

Visualisasi ruang *embedding* dilakukan untuk memahami bagaimana struktur semantik judul penelitian berubah setelah proses *whitening*. Tiga teknik reduksi dimensi digunakan untuk mengevaluasi karakteristik lokal, global, dan pola kesamaan *embedding*. Setiap teknik memberikan sudut pandang yang berbeda sehingga menghasilkan pemahaman yang lebih komprehensif mengenai efektivitas *Ditto Whitening* dalam memperbaiki kualitas *embedding*.

1) t-SNE

t-SNE (*t-distributed Stochastic Neighbor Embedding*) digunakan untuk memvisualisasikan kedekatan lokal *embedding*, yaitu bagaimana judul-judul penelitian yang semestinya mirip secara semantik dikelompokkan dalam ruang dua dimensi[21]. Metode ini memfokuskan pada pelestarian probabilitas kedekatan antar-titik pada ruang asli sehingga menghasilkan *cluster* padat untuk kelompok judul yang memiliki representasi serupa.

2) UMAP

UMAP (*Uniform Manifold Approximation and Projection*) digunakan untuk mengobservasi struktur *manifold global embedding*[22]. Berbeda dengan t-SNE yang lebih sensitif terhadap hubungan lokal, UMAP mampu mempertahankan hubungan struktur secara keseluruhan dan memberikan gambaran yang lebih stabil mengenai distribusi *embedding* secara global.

3) *Heatmap Cosine Similarity*

Heatmap cosine similarity digunakan untuk memeriksa pola kemiripan antarjudul penelitian, mendeteksi bias *cosine*, serta mengidentifikasi sejauh mana distribusi *embedding* berubah sebelum dan sesudah *whitening*. Pada IndoBERT *baseline*, heatmap memperlihatkan pola blok dan garis diagonal tebal, menandakan adanya nilai *cosine* tinggi terhadap banyak pasangan judul, suatu ciri dari *embedding* anisotropik yang memusat pada arah tertentu.

4) Histogram Distribusi *Cosine Similarity*

Histogram distribusi *cosine similarity* digunakan untuk mengevaluasi pola persebaran kedekatan antarjudul penelitian dalam ruang *embedding*. Tidak seperti heatmap yang menampilkan hubungan antarsemua pasangan judul dalam bentuk matriks, histogram memberikan gambaran agregat mengenai seberapa sering nilai *cosine* tertentu muncul, sehingga memudahkan analisis pergeseran distribusi secara menyeluruh. Teknik ini sering dipakai untuk mendeteksi fenomena *anisotropic collapse* pada model berbasis transformer[4].

III. HASIL DAN PEMBAHASAN

A. Hasil Evaluasi Kualitas Embedding

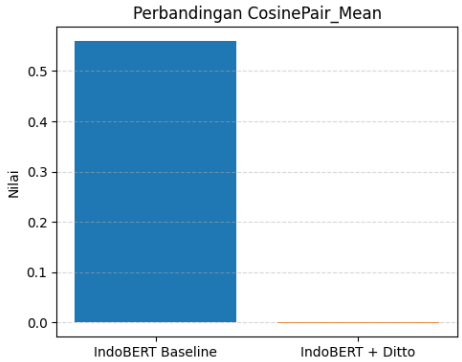
Hasil Evaluasi kualitas *embedding* dilakukan dengan membandingkan dua konfigurasi utama, yaitu IndoBERT + *L2 normalization* sebagai *baseline* dan IndoBERT + *Ditto Whitening* sebagai model yang diusulkan. Evaluasi dilakukan menggunakan metrik intrinsik yang mencerminkan karakteristik geometri ruang *embedding*, meliputi *Isotropy Index, Cosine Pair Mean, MeanCos-to-Mean, Uniformity Loss,* dan *Hubness Skew*. Hasil perbandingan kuantitatif dari seluruh metrik evaluasi intrinsik pada konfigurasi *baseline* dan model yang diusulkan dirangkum dalam Tabel 4.

TABEL 4. METRIK KUALITAS *EMBEDDING*

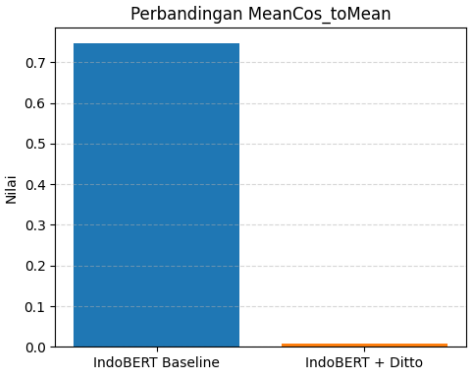
Metrik Evaluasi	IndoBERT Baseline	IndoBERT + Ditto Whitening	Perubahan / Implikasi
IsotropyIndex (Eigenvalue)	2.0794×10^{-10}	4.6263×10^{-6}	↑ 22.000× lebih isotropik
CosinePair Mean	0.5597	-0.000145	Nilai *cosine* acak mendekati 0 (lebih isotropik)
MeanCos_to Mean	0.747	0.0068	Bias *mean* menghilang drastis
Uniformity Loss	-1.6754	-3.9849	*Embedding* lebih *uniform* pada *manifold*
Hubness Skew	1.6008	0.6804	Fenomena *hubness* berkurang signifikan

Tabel 4 menyajikan perbandingan kuantitatif seluruh metrik evaluasi antara konfigurasi *baseline* dan *embedding* yang telah dioptimalkan menggunakan *Ditto Whitening*. Secara umum, seluruh metrik menunjukkan perbaikan yang konsisten setelah penerapan *whitening*.

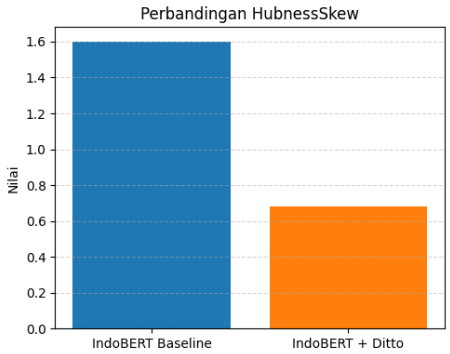
Nilai *Isotropy Index* meningkat secara signifikan, yang menunjukkan bahwa distribusi variansi antar dimensi *embedding* menjadi lebih merata. Pada saat yang sama, penurunan *Cosine Pair Mean* dan *MeanCos-to-Mean* mengindikasikan berkurangnya bias kemiripan global, di mana *embedding* tidak lagi terkonsentrasi pada satu arah dominan. Selain itu, penurunan nilai *Hubness Skew* menunjukkan berkurangnya fenomena *hubness*, sehingga ruang *embedding* menjadi lebih seimbang untuk kebutuhan pencarian tetangga terdekat. Visualisasi perbandingan metrik tersebut disajikan pada Gambar 3–5.



Gambar 3. Perbandingan *CosinePair Mean*



Gambar 4. Perbandingan *MeanCos-to-Mean*

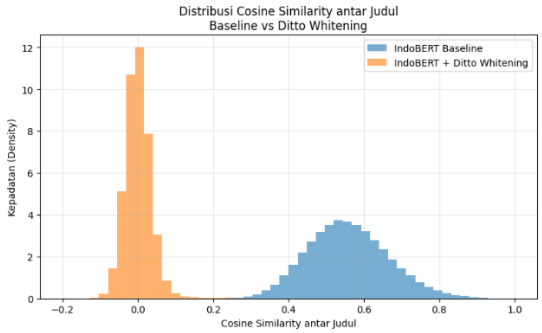


Gambar 5. Perbandingan *Hubness Skew*

Ketiga visualisasi tersebut memperkuat hasil kuantitatif pada Tabel 4 dengan menunjukkan perubahan distribusi nilai *cosine* dan tingkat *hubness* sebelum dan sesudah penerapan *Ditto Whitening*.

B. Distribusi Cosine dan Pemerataan Ruang Embedding

Distribusi nilai *cosine similarity* antar judul penelitian menunjukkan perbedaan yang jelas antara *embedding baseline* dan *embedding* hasil *whitening*. Pada konfigurasi *baseline*, sebagian besar pasangan judul memiliki nilai *cosine* yang relatif tinggi, yang mengindikasikan adanya bias kemiripan global akibat anisotropi. Setelah *Ditto Whitening* diterapkan, distribusi nilai *cosine* menjadi lebih terpusat di sekitar nol dan lebih simetris, mencerminkan peningkatan isotropi ruang *embedding,* sebagaiamana divisualisasikan pada Gambar 6.



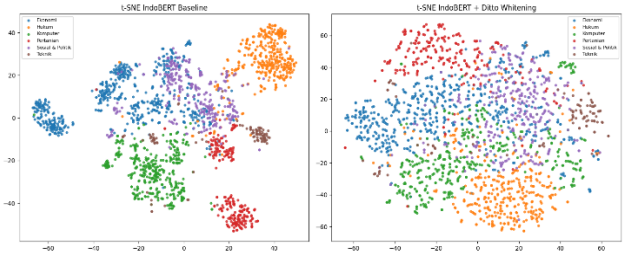
Gambar 6. Distribusi *Cosine Similarity* antar Judul

Perubahan distribusi ini menunjukkan bahwa *embedding* hasil *whitening* lebih mampu merepresentasikan perbedaan semantik antar judul penelitian, tanpa terdorong oleh kesamaan struktural atau terminologi umum semata.

Penurunan nilai *Uniformity Loss* pada Tabel 4. menunjukkan bahwa *embedding* menyebar lebih merata pada *manifold*-nya, sehingga tidak lagi terkonsentrasi pada wilayah tertentu dalam ruang *embedding*.

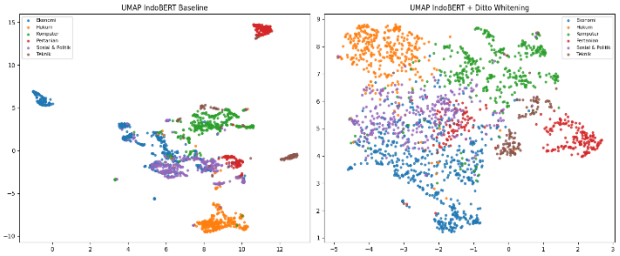
C. Interpretasi Visual Ruang Embedding

Visualisasi ruang *embedding* menggunakan teknik reduksi dua dimensi, kemudian dilakukan untuk memperjelas perubahan struktur ruang semantik sebelum dan sesudah penerapan *Ditto Whitening*, sebagaimana ditunjukkan pada Gambar 7.



Gambar 7. Visualisasi t-SNE IndoBERT *Baseline* vs IndoBERT+Ditto

Pada *embedding baseline*, titik-titik judul penelitian terlihat menggumpal dan sulit dibedakan antar rumpun ilmu. Setelah penerapan *Ditto Whitening*, penyebaran titik menjadi lebih merata dan struktur klaster terlihat lebih terorganisasi, sebagaimana ditunjukkan pada Gambar 8.

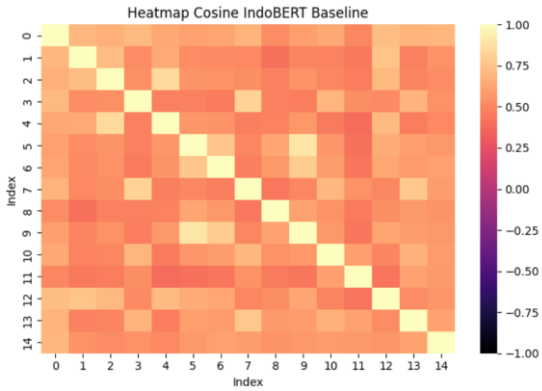


Gambar 8. Visualisasi UMAP IndoBERT *Baseline* vs IndoBERT+Ditto

Beberapa rumpun ilmu, seperti Ekonomi dan Sosial Politik, masih tampak berdekatan pada visualisasi dua dimensi. Kedekatan ini dapat dijelaskan oleh adanya tumpang tindih semantik pada judul-judul penelitian yang membahas tema sosial, kebijakan publik, dan perilaku masyarakat dengan terminologi yang serupa. Oleh karena itu, kedekatan klaster tersebut mencerminkan kemiripan semantik domain, bukan kegagalan representasi *embedding*.

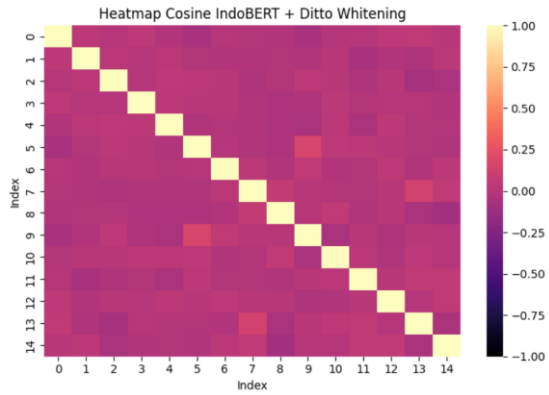
D. Analisis Heatmap Cosine Similarity

Heatmap cosine similarity digunakan untuk mengevaluasi pola kemiripan antarjudul penelitian secara lebih rinci, sebagaimana ditunjukkan pada Gambar 9.



Gambar 9. *Heatmap Cosine* IndoBERT *Baseline*

Pada *embedding baseline*, *heatmap* menunjukkan pola warna terang di luar diagonal utama, yang menandakan banyak pasangan judul memiliki nilai *cosine* tinggi meskipun tidak berkaitan secara semantic, sebagaimana ditunjukkan pada Gambar 10.



Gambar 10. *Heatmap Cosine* IndoBERT + *Ditto Whitening*

Setelah *Ditto Whitening*, *heatmap* didominasi oleh warna gelap di luar diagonal utama, yang menunjukkan bahwa bias kemiripan global telah berkurang secara signifikan. Hanya diagonal utama yang mempertahankan nilai *cosine* tinggi, sesuai dengan kemiripan diri sendiri.

E. Pembahasan Teoretis dan Implikasi

Hasil penelitian ini sejalan dengan teori *anisotropic collapse* pada model transformer yang dikemukakan oleh Ethayarajh, yang menjelaskan bahwa *embedding* berbasis BERT cenderung terkonsentrasi pada subruang tertentu sehingga menghasilkan bias kemiripan global dan nilai *cosine similarity* yang tinggi antar pasangan teks yang tidak berkaitan secara semantik[4]. Kondisi tersebut juga memicu fenomena hubness, di mana sebagian vektor menjadi tetangga terdekat bagi banyak vektor lain secara tidak proporsional. Penurunan nilai *Hubness Skew* setelah penerapan *Ditto Whitening* pada penelitian ini menunjukkan bahwa distribusi *embedding* menjadi lebih seimbang,

sehingga meningkatkan keandalan sistem dalam tugas *nearest neighbor search*, yang merupakan komponen inti dalam sistem pencarian dan deteksi kemiripan judul penelitian.

Temuan ini konsisten dengan penelitian sebelumnya yang melaporkan bahwa metode *whitening* dan normalisasi *embedding* mampu meningkatkan isotropi dan kualitas representasi semantik. Su et al. menunjukkan bahwa pemerataan variansi antar dimensi *embedding* melalui proses *whitening* dapat mengurangi bias distribusi dan meningkatkan performa pengukuran kesamaan semantik[6]. Lebih lanjut, Li et al. memperkenalkan *Ditto Whitening* sebagai pendekatan transformasi *linear* berbasis kovarians yang dirancang untuk meningkatkan isotropi *embedding* tanpa merusak struktur semantik internal[10]. Keunggulan *Ditto Whitening* dibandingkan metode *whitening* lain terletak pada kemampuannya mempertahankan relasi relatif antar vektor tanpa memerlukan pelatihan tambahan atau asumsi distribusi yang kompleks, sehingga menjadikannya sangat sesuai untuk optimasi *embedding* pada konteks teks akademik pendek berbahasa Indonesia.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan *Ditto Whitening* secara efektif meningkatkan kualitas *embedding* IndoBERT pada konteks judul penelitian berbahasa Indonesia melalui peningkatan isotropi, pengurangan bias kemiripan global, dan penurunan fenomena *hubness* pada ruang *embedding*. Perbaikan karakteristik geometris tersebut meningkatkan keandalan pengukuran kesamaan semantik, sehingga mendukung pengembangan sistem deteksi kemiripan judul dan pencarian dokumen akademik yang lebih akurat. Meskipun demikian, penelitian ini memiliki keterbatasan karena evaluasi dilakukan secara intrinsik, menggunakan satu varian model bahasa dan dataset yang bersumber dari satu institusi perguruan tinggi. Oleh karena itu, penelitian lanjutan disarankan untuk mengevaluasi efektivitas *Ditto Whitening* pada model *embedding* yang lebih mutakhir, seperti E5 dan BGE-M3, serta pada dataset lintas institusi. Secara praktis, temuan ini berpotensi diintegrasikan ke dalam sistem repositori dan layanan manajemen topik penelitian institusional sebagai bagian dari penguatan tata kelola akademik dan transformasi digital di lingkungan pendidikan tinggi.

REFFERENSI

[1] P. Zhang, X. Huang, Y. Wang, C. Jiang, and S. Member, "Semantic Similarity Computing Model Based on Multi Model Fine-Grained Nonlinear Fusion," vol. 9, 2021, doi: 10.1109/ACCESS.2021.3049378.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019, doi: https://doi.org/10.18653/v1/N19-1423.

[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 3982–3992, 2019, doi: 10.18653/v1/d19-1410.

[4] K. Ethayarajh, "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMO, and GPT-2 Embeddings," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 55–65, 2019, doi: 10.18653/v1/d19-1006.

[5] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the Sentence Embeddings from Pre-trained Language Models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9119–9130, doi: 10.18653/v1/2020.emnlp-main.733.

[6] J. Su, J. Cao, W. Liu, and Y. Ou, "Whitening Sentence Representations for Better Semantics and Faster Retrieval," 2021, doi: https://doi.org/10.48550/arXiv.2103.15316.

[7] J. Mu and P. Viswanath, "All-but-the-top: Simple and Effective Post-Processing for Word Representations," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–25, 2018, doi: https://doi.org/10.48550/arXiv.1702.01417.

[8] T. Jiang *et al.*, "PromptBERT: Improving BERT Sentence Embeddings with Prompts," *arXiv Prepr. arXiv2201.04337*, 2022, doi: https://doi.org/10.48550/arXiv.2201.04337.

[9] Jumino and S. A. Suwanto, "Analisis Layanan Repositori Universitas Diponegoro Berdasarkan Aksesibilitas, Tampilan, Dan Isi: Upaya Pemberdayaan Repositori Berbasis Riset," vol. 9008, no. 21, 2019, doi: 10.14203/j.baca.v40i2.449.

[10] Q. Chen *et al.*, "Ditto: A Simple and Efficient Approach to Improve Sentence Embeddings," *EMNLP 2023 - 2023 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 5868–5875, 2023, doi: 10.18653/v1/2023.emnlp-main.359.

[11] BAAK, "Sistem Informasi Akademik Unisan," 2025. https://siakun.unisan.ac.id/ (accessed Aug. 25, 2025).

[12] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," no. 2.

[13] I. Garrido-muñoz, A. Montejo-ráez, F. Martínez-santiago, and L. A. Ureña-lópez, "A Survey on Bias in Deep NLP," 2021, doi: https://doi.org/10.3390/app11073184.

[14] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[15] I. T. Jolliffe, *Principal component analysis*, 2nd ed. Springer, 2002.

[16] A. Kessy, A. Lewin, and K. Strimmer, "Optimal Whitening and Decorrelation," no. December 2015, pp. 1–14, 2016, doi: https://doi.org/10.1080/00031305.2016.1277159.

[17] C. M. Bishop and N. M. Nasrabadi, *Pattern*

Recognition and Machine Learning, vol. 4, no. 4. Springer, 2006.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 1. MIT press Cambridge, 2016.

[19] T. Wang and P. Isola, "Understanding Contrastive Representation Learning Through Alignment and Uniformity on the Hypersphere," *37th Int. Conf. Mach. Learn. ICML 2020*, vol. PartF16814, pp. 9871–9881, 2020.

[20] R. Feldbauer, T. Rattei, and A. Flexer, "scikit-hubness : Hubness Reduction and Approximate Neighbor Search," vol. 5, pp. 45–47, 2020, doi: 10.21105/joss.01957.

[21] L. Van Der Maaten and G. Hinton, "Visualizing Data Using t-SNE," vol. 9, pp. 2579–2605, 2008.

[22] L. Mcinnes, J. Healy, and J. Melville, "UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction," 2020, doi: https://doi.org/10.48550/arXiv.1802.03426.