Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 3, 2025
DOI: 10.30829/zero.v9i3.26549
Page: 1073 - 1084

Evaluating Semantic Geometry of Indonesian News Texts: Agglomerative Clustering Study using IndoBERT Embeddings

[1] Joni Wilson Sitopu
Faculty of Engineering, Universitas Simalungun, Pematangsiantar, Indonesia

Article Info

Article history:

Accepted 26 December 2025

Keywords:

Agglomerative Clustering;
Euclidean Distance;
IndoBERT Embeddings;
Indonesian News Analysis;
Semantic Geometry;
UMAP

ABSTRACT

This study evaluates the effectiveness of various Agglomerative Clustering configurations in uncovering the Semantic Geometry of Indonesian online news texts represented by IndoBERT embeddings. Using a comparative experimental design, the research assessed clustering performance based on different distance metrics and dimensionality reduction methods. Results show that Cosine Similarity produced a highly skewed distribution, with over 99% of documents (35,766) concentrated in one cluster, indicating limited thematic differentiation. In contrast, the UMAP–Euclidean configuration achieved the most balanced distribution (4,254–8,204 documents per cluster) and the highest topic coherence score of 0.71, outperforming previous IndoBERT clustering studies using K-Means (0.54) or DBSCAN (0.59). Thematically, five major domains were identified: Politics, Health & Technology, Macroeconomics & Finance, Economy & Industry, and Education & Social Issues. These findings demonstrate that non-linear UMAP projection significantly enhances thematic granularity, offering methodological advancement and practical benefits for computational linguistics, semantic text analysis in Indonesian language research.

Corresponding Author:

Joni Wilson Sitopu,
Faculty of Engineering
Universitas Simalungun, Pematangsiantar, Indonesia
Email: jwsitopu@gmail.com

1. INTRODUCTION

Indonesian online news texts constitute a vast and dynamic corpus that mirrors the evolving landscape of national discourse from politics and economics to social and technological developments [1], [2], [3]. However, linguistic diversity in authorship, regional expressions, and temporal variations makes automated text interpretation particularly challenging [4], [5], [6]. Two semantically equivalent news statements such as “harga pangan naik” [food prices are rising] and “biaya makan meningkat” [cost of food is increasing] illustrate how lexical diversity can obscure shared meaning. Conventional lexical-based methods, including Bag-of-Words and TF-IDF, are limited in capturing these semantic nuances since they rely solely on surface-level token frequency [7], [8], [9].

To address this limitation, recent developments have focused on semantic representation using transformer-based models such as IndoBERT, which embed sentences into high-dimensional vectors that encode both linguistic context and meaning [10], [11], [12]. This vector space often conceptualized as Semantic Geometry enables spatial computation of meaning through measures such as Cosine Similarity, allowing more robust thematic mapping of la rge textual datasets [13], [14], [15]. However, despite improvements in contextual understanding, clustering within such dense embedding spaces remains a methodological challenge, as semantically similar documents often overlap in representation [16], [17], [18].

Previous comparative analyses reveal that lexical-based clustering models (e.g., TF-IDF + K-Means) typically achieve average semantic coherence scores between 0.42 and 0.51 on Indonesian corpora [7], while early transformer-based approaches like IndoBERT without dimensionality optimization reach around 0.63 [11]. Yet, thematic separability remains limited, with intra-cluster similarity exceeding 0.78 across semantically distinct topics [12]. These findings suggest the need for an integrated approach that not only leverages semantic embeddings but also optimizes cluster distinction through adaptive distance metrics and dimensionality reduction. This study thus seeks to empirically evaluate how variations in distance metrics (Cosine vs. Euclidean) and reduction techniques (PCA, UMAP) influence the clarity of topic boundaries within the IndoBERT embedding space.

Building on these insights, the present study aims to construct a semantic clustering framework for Indonesian news data that achieves higher thematic precision and interpretability. Accordingly, the research is guided by the following problem formulation: Which combination of distance metrics and dimensionality reduction techniques yields the most coherent and thematically distinct clustering of Indonesian online news embeddings generated by IndoBERT? By addressing this question, the study contributes to the methodological refinement of semantic modeling for low-resource languages and provides a scalable framework for automated topic detection in Indonesian media analysis.

2. RESEARCH METHOD

The following section systematically outlines the methodological stages employed in this study to ensure that the data analysis process is structured, transparent, and reproducible.

2.1 General Architecture

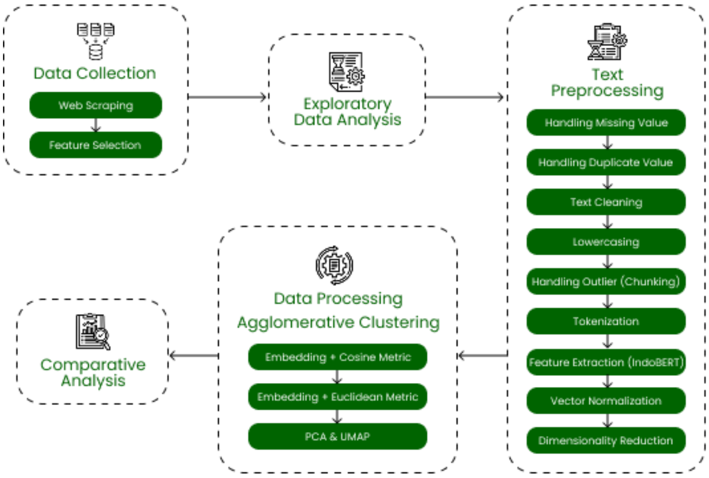
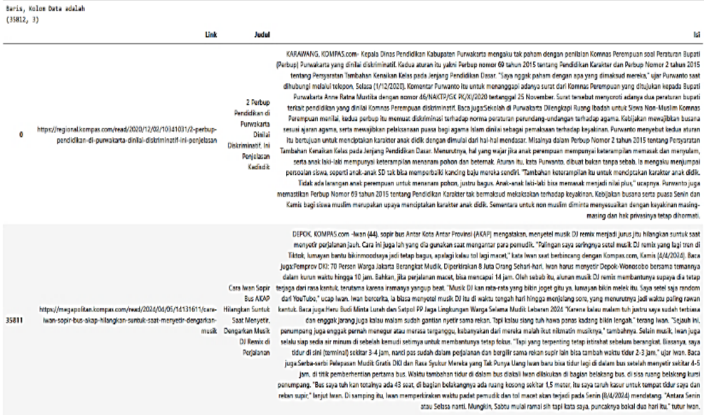


Figure 1. General Architecture

2.2 Data Collection

The data in this study were collected through web scraping from the online news website Kompas News (https://www.kompas.com) using the requests and BeautifulSoup libraries. The scraping process was carried out in two stages: first, extracting all article links based on the categories politics, economy, health, sports, education, technology, and entertainment, and then retrieving the news content from those links. As a result, a total of 35,854 news links were gathered, with 35,812 articles successfully extracted and stored in CSV format.



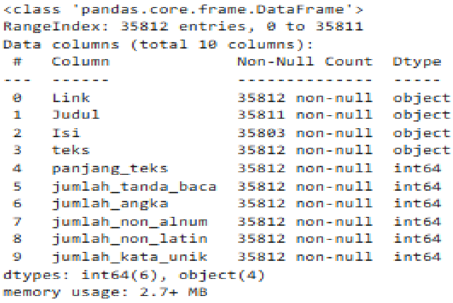
```
Baris, Kolom Data adalah
(35812, 3)
```

	Link	Judul	Isi
0	https://regional.kompas.com/read/2020/12/02/10341031/2-perbup-pendidikan-di-purwakarta-dinilai-diskriminatif-ini-penjelasan	2 Perbup Pendidikan di Purwakarta Dinilai Diskriminatif, Ini Penjelasan Kadisdik	KARAWANG, KOMPAS.com- Kepala Dinas Pendidikan Kabupaten Purwakarta mengaku tak paham dengan penilaian Komnas Perempuan soal Peraturan Bupati (Perbup) Purwakarta yang dinilai diskriminatif. Kedua aturan itu yakni Perbup nomor 69 tahun 2015 tentang Pendidikan Karakter dan Perbup Nomor 2 tahun 2015 tentang Persyaratan Tambahan Kenaikan Kelas pada Jenjang Pendidikan Dasar. "Saya nggak paham dengan apa yang dimaksud mereka," ujar Purwanto saat dihubungi melalui telepon, Selasa (1/12/2020). Komentar Purwanto itu untuk menanggapi adanya surat dari Komnas Perempuan yang ditujukan kepada Bupati Purwakarta Anne Ratna Mustika dengan nomor 46/NAKTP/GK PK/XI/2020 tertanggal 25 November. Surat tersebut menyoroti adanya dua peraturan bupati terkait pendidikan yang dinilai Komnas Perempuan diskriminatif. Baca juga:Sekolah di Purwakarta Dilengkapi Ruang Ibadah untuk Siswa Non-Muslim Komnas Perempuan menilai, kedua perbup itu memuat diskriminasi terhadap norma peraturan perundang-undangan terhadap agama. Kebijakan mewajibkan busana sesuai ajaran agama, serta mewajibkan pelaksanaan puasa bagi agama Islam dinilai sebagai pemaksaan terhadap keyakinan. Purwanto menyebut kedua aturan itu bertujuan untuk menciptakan karakter anak didik dengan dimulai dari hal-hal mendasar. Misalnya dalam Perbup Nomor 2 tahun 2015 tentang Persyaratan Tambahan Kenaikan Kelas pada Jenjang Pendidikan Dasar. Menurutnya, hal yang wajar jika anak perempuan mempunyai keterampilan memasak dan menyulam, serta anak laki-laki mempunyai keterampilan menanam pohon dan beternak. Aturan itu, kata Purwanto, dibuat bukan tanpa sebab. Ia mengaku menjumpai persoalan siswa, seperti anak-anak SD tak bisa memperbaiki kancing baju mereka sendiri. "Tambahan keterampilan itu untuk menciptakan karakter anak didik. Tidak ada larangan anak perempuan untuk menanam pohon, justru bagus. Anak-anak laki-laki bisa memasak menjadi nilai plus," ucapnya. Purwanto juga memastikan Perbup Nomor 69 tahun 2015 tentang Pendidikan Karakter tak bermaksud memaksakan terhadap keyakinan. Kebijakan busana serta puasa Senin dan Kamis bagi siswa muslim merupakan upaya menciptakan karakter anak didik. Sementara untuk non muslim diminta menyesuaikan dengan keyakinan masing-masing dan hak privasinya tetap dihormati.
35811	https://megapolitan.kompas.com/read/2024/04/05/14131611/cara-iwan-sopir-bus-akap-hilangkan-suntuk-saat-menyetir-dengarkan-musik	Cara Iwan Sopir Bus AKAP Hilangkan Suntuk Saat Menyetir, Dengarkan Musik DJ Remix di Perjalanan	DEPOK, KOMPAS.com -Iwan (44), sopir bus Antar Kota Antar Provinsi (AKAP) mengatakan, menyetel musik DJ remix menjadi jurus jitu hilangkan suntuk saat menyetir perjalanan jauh. Cara ini juga lah yang dia gunakan saat mengantar para pemudik. "Palingan saya seringnya setel musik DJ remix yang lagi tren di Tiktok, lumayan bantu bikinmoodsaya jadi tetap bagus, apalagi kalau tol lagi macet," kata Iwan saat berbincang dengan Kompas.com, Kamis (4/4/2024). Baca juga:Pemprov DKI: 70 Persen Warga Jakarta Berangkat Mudik, Diperkirakan 8 Juta Orang Sehari-hari. Iwan harus menyetir Depok-Wonosobo bersama temannya dalam kurun waktu hingga 10 jam. Bahkan, jika perjalanan macet, bisa mencapai 14 jam. Oleh sebab itu, alunan musik DJ remix membantunya supaya dia tetap terjaga dari rasa kantuk, terutama karena iramanya yangup beat. "Musik DJ kan rata-rata yang bikin joget gitu ya, lumayan bikin melek itu. Saya setel saja random dari YouTube," ucap Iwan. Iwan bercerita, ia biasa menyetel musik DJ itu di waktu tengah hari hingga menjelang sore, yang menurutnya jadi waktu paling rawan kantuk. Baca juga:Heru Budi Minta Lurah dan Satpol PP Jaga Lingkungan Warga Selama Mudik Lebaran 2024 "Karena kalau malam tuh justru saya sudah terbiasa dan enggak jarang juga kalau malam sudah gantian nyetir sama rekan. Tapi kalau siang tuh hawa panas kadang bikin lengah," terang Iwan. "Sejauh ini, penumpang juga enggak pernah menegur atau merasa terganggu, kebanyakan dari mereka malah ikut nikmatin musiknya," tambahnya. Selain musik, Iwan juga selalu siap sedia air minum di sebelah kemudi setirnya untuk membantunya tetap fokus. "Tapi yang terpenting tetap istirahat sebelum berangkat. Biasanya, saya tidur di sini (terminal) sekitar 3-4 jam, nanti pas sudah dalam perjalanan dan bergilir sama rekan supir lain bisa tambah waktu tidur 2-3 jam," ujar Iwan. Baca juga:Serba-serbi Pelepasan Mudik Gratis DKI dan Rasa Syukur Mereka yang Tak Punya Uang Iwan baru bisa tidur lagi di dalam bus setelah menyetir sekitar 4-5 jam, di titik pemberhentian pertama bus. Waktu tambahan tidur di dalam bus diakali Iwan dilakukan di bagian belakang bus, di sisa ruang belakang kursi penumpang. "Bus saya tuh kan totalnya ada 43 seat, di bagian belakangnya ada ruang kosong sekitar 1,5 meter, itu saya taruh kasur untuk tempat tidur saya dan rekan supir," lanjut Iwan. Di samping itu, Iwan memperkirakan waktu padat pemudik dan tol macet akan terjadi pada Senin (8/4/2024) mendatang. "Antara Senin atau Selasa nanti. Mungkin, Sabtu mulai ramai sih tapi kata saya, puncaknya bakal dua hari itu," tutur Iwan.

Figure 2. Sample of The Scraped Data

2.3 Exploratory Data Analysis (EDA)

Data exploration was carried out to understand the characteristics of the dataset and prepare it before the processing stage. The analysis included checking for missing values, duplicates, text length distribution, outliers, and unusual characters. The results showed the presence of missing values, a total of 35,812 documents with an average text length of 2,654 characters, 29 lower outliers and 1,192 upper outliers, as well as noisy elements such as “baca juga,” “editor kompas.com,” and emojis that needed to be cleaned during preprocessing.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35812 entries, 0 to 35811
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Link                35812 non-null  object
 1   Judul               35811 non-null  object
 2   Isi                 35803 non-null  object
 3   teks                35812 non-null  object
 4   panjang_teks        35812 non-null  int64
 5   jumlah_tanda_baca   35812 non-null  int64
 6   jumlah_angka        35812 non-null  int64
 7   jumlah_non_alnum    35812 non-null  int64
 8   jumlah_non_latin    35812 non-null  int64
 9   jumlah_kata_unik    35812 non-null  int64
dtypes: int64(6), object(4)
memory usage: 2.7+ MB
```

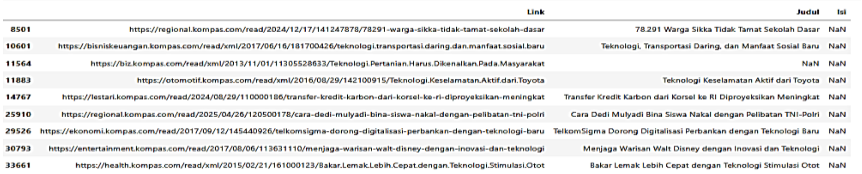
Figure 3. The Data Info

2.4 Text Preprocessing

The text preprocessing stage was systematically conducted to prepare the raw dataset before semantic modeling. The process began with handling missing and duplicate values to ensure data integrity. Missing entries in the Title and Content columns were removed because they contained non-textual content such as videos. These missing values were detected using a Boolean mask and subsequently deleted from the dataset. Duplicate checking was also performed using hash-based equality verification, and no duplicate data were found.

Next, text cleaning was performed using the clean_text() function to eliminate irrelevant components, including HTML tags, media phrases (“kompas.com”, “baca juga”), symbols, numbers, and non-Latin characters. Spaces were also normalized using Regex to ensure uniform text formatting. All text was then converted to lowercase to maintain lexical consistency during tokenization and embedding processes.

To handle long documents, a chunking process was applied to divide texts exceeding 512 tokens according to IndoBERT’s tokenization limit. The embeddings of all chunks were then aggregated using mean embedding, as it provides a stable and efficient representation that captures the overall semantic content of long documents while maintaining consistent vector dimensionality for reliable clustering. Tokenization was performed using IndoBERT’s AutoTokenizer based on the IndoBERT vocabulary.

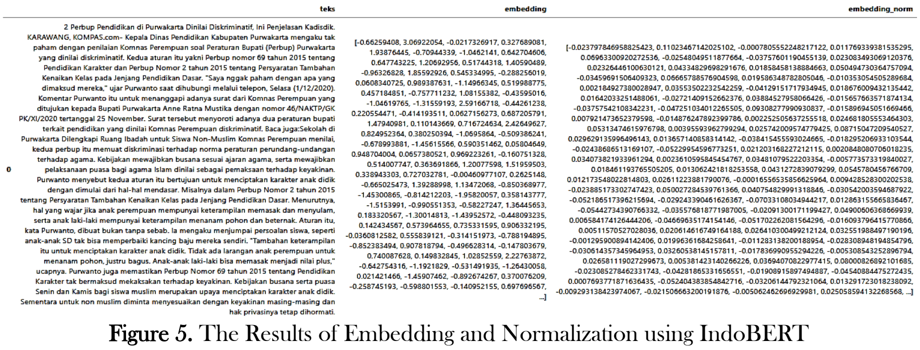


	Link	Judul	Isi
8501	https://regional.kompas.com/read/2024/12/17/141247878/78291-warga-sikka-tidak-tamat-sekolah-dasar	78.291 Warga Sikka Tidak Tamat Sekolah Dasar	NaN
10601	https://bisniskeuangan.kompas.com/read/xml/2017/06/16/181700426/teknologi.transportasi.daring.dan.manfaat.sosial.baru	Teknologi, Transportasi Daring, dan Manfaat Sosial Baru	NaN
11564	https://biz.kompas.com/read/xml/2013/11/01/11305528633/Teknologi.Pertanian.Harus.Dikenalkan.Pada.Masyarakat	NaN	NaN
11883	https://otomotif.kompas.com/read/xml/2016/08/29/142100915/Teknologi.Keselamatan.Aktif.dari.Toyota	Teknologi Keselamatan Aktif dari Toyota	NaN
14767	https://lestari.kompas.com/read/2024/08/29/110000186/transfer-kredit-karbon-dari-korsel-ke-ri-diproyeksikan-meningkat	Transfer Kredit Karbon dari Korsel ke RI Diproyeksikan Meningkat	NaN
25910	https://regional.kompas.com/read/2025/04/26/120500178/cara-dedi-mulyadi-bina-siswa-nakal-dengan-pelibatan-tni-polri	Cara Dedi Mulyadi Bina Siswa Nakal dengan Pelibatan TNI-Polri	NaN
29526	https://ekonomi.kompas.com/read/2017/09/12/145440926/telkomsigma-dorong-digitalisasi-perbankan-dengan-teknologi-baru	TelkomSigma Dorong Digitalisasi Perbankan dengan Teknologi Baru	NaN
30793	https://entertainment.kompas.com/read/2017/08/06/113631110/menjaga-warisan-walt-disney-dengan-inovasi-dan-teknologi	Menjaga Warisan Walt Disney dengan Inovasi dan Teknologi	NaN
33661	https://health.kompas.com/read/xml/2015/02/21/161000123/Bakar.Lemak.Lebih.Cepat.dengan.Teknologi.Stimulasi.Otot	Bakar Lemak Lebih Cepat dengan Teknologi Stimulasi Otot	NaN

Figure 4. Data with Missing Values

During feature extraction, each chunk was transformed into a 768-dimensional semantic vector using the transformer encoder architecture. For longer texts, the mean vector value across all chunks was calculated to produce a unified document representation. To ensure consistency, L2-normalization was applied so that all vectors had uniform magnitude, thereby stabilizing similarity measurements during Cosine Similarity computation.

Finally, dimensionality reduction was performed in two sequential stages. The PCA (Principal Component Analysis) method reduced the 768-dimensional embeddings to 50 dimensions linearly, minimizing noise while retaining essential variance. Subsequently, UMAP (Uniform Manifold Approximation and Projection) further reduced the data to two dimensions using the Cosine metric, preserving local semantic relationships and enabling clear visualization of thematic clusters in a two-dimensional semantic space:



	teks	embedding	embedding_norm
0	2 Perbup Pendidikan di Purwakarta Dinilai Diskriminatif, Ini Penjelasan Kadisdik. KARAWANG, KOMPAS.com- Kepala Dinas Pendidikan Kabupaten Purwakarta mengaku tak paham dengan penilaian Komnas Perempuan soal Peraturan Bupati (Perbup) Purwakarta yang dinilai diskriminatif. Kedua aturan itu yakni Perbup nomor 69 tahun 2015 tentang Pendidikan Karakter dan Perbup Nomor 2 tahun 2015 tentang Persyaratan Tambahan Kenaikan Kelas pada Jenjang Pendidikan Dasar. "Saya nggak paham dengan apa yang dimaksud mereka," ujar Purwanto saat dihubungi melalui telepon, Selasa (1/12/2020). Komentar Purwanto itu untuk menanggapi adanya surat dari Komnas Perempuan yang ditujukan kepada Bupati Purwakarta Anne Ratna Mustika dengan nomor 46/NAKTP/GK PK/XI/2020 tertanggal 25 November. Surat tersebut menyoroti adanya dua peraturan bupati terkait pendidikan yang dinilai Komnas Perempuan diskriminatif. Baca juga:Sekolah di Purwakarta Dilengkapi Ruang Ibadah untuk Siswa Non-Muslim Komnas Perempuan menilai, kedua perbup itu memuat diskriminasi terhadap norma peraturan perundang-undangan terhadap agama. Kebijakan mewajibkan busana sesuai ajaran agama, serta mewajibkan pelaksanaan puasa bagi agama Islam dinilai sebagai pemaksaan terhadap keyakinan. Purwanto menyebut kedua aturan itu bertujuan untuk menciptakan karakter anak didik dengan dimulai dari hal-hal mendasar. Misalnya dalam Perbup Nomor 2 tahun 2015 tentang Persyaratan Tambahan Kenaikan Kelas pada Jenjang Pendidikan Dasar. Menurutnya, hal yang wajar jika anak perempuan mempunyai keterampilan memasak dan menyulam, serta anak laki-laki mempunyai keterampilan menanam pohon dan beternak. Aturan itu, kata Purwanto, dibuat bukan tanpa sebab. Ia mengaku menjumpai persoalan siswa, seperti anak-anak SD tak bisa memperbaiki kancing baju mereka sendiri. "Tambahan keterampilan itu untuk menciptakan karakter anak didik. Tidak ada larangan anak perempuan untuk menanam pohon, justru bagus. Anak-anak laki-laki bisa memasak menjadi nilai plus," ucapnya. Purwanto juga memastikan Perbup Nomor 69 tahun 2015 tentang Pendidikan Karakter tak bermaksud mekaksakan terhadap keyakinan. Kebijakan busana serta puasa Senin dan Kamis bagi siswa muslim merupakan upaya menciptakan karakter anak didik. Sementara untuk non muslim diminta menyesuaikan dengan keyakinan masing-masing dan hak privasinya tetap dihormati.	[-0.66259408, 3.06922054, -0.0217326917, 0.327689081, 1.93876445, -0.70944339, -1.0462141, 0.642704606, 0.647743225, 1.20692956, 0.51744318, 1.40590489, -0.96326828, 1.85592926, 0.545334995, -0.288256019, 0.0608340725, 0.989387631, -1.14966345, 0.519988775, 0.457184851, -0.757711232, 1.08155382, -0.43595016, -1.04619765, -1.31559193, 2.59166718, -0.44261238, 0.220554471, -0.414193511, 0.0627156273, 0.687205791, 1.47940981, 0.110143669, 0.716724634, 2.42649627, 0.824952364, 0.380250394, -1.0695864, -0.509386241, -0.678993881, -1.45615566, 0.590351462, 0.05804649, 0.948704004, 0.0657380521, 0.969223261, -0.160751328, 0.514007747, 0.363691866, 1.20077598, 1.51959503, 0.338943303, 0.727032781, -0.00460977107, 0.2625148, -0.665025473, 1.39288998, 1.13472068, -0.850368977, -1.45300865, -0.814212203, -1.95820057, 0.358143777, -1.5153991, -0.990551353, -0.58227247, 1.36445653, 0.183320567, -1.30014813, -1.43952572, -0.448093235, 0.142434567, 0.573964655, 0.735331595, 0.906332195, -0.0360812582, 0.555839121, -0.314151973, -0.788194895, -0.852383494, 0.907818794, -0.496628314, -0.147803679, 0.740087628, 0.149832845, 1.02852559, 2.22763872, -0.642754316, -1.1921829, -0.531491935, -1.26430058, 0.021421466, -1.45907462, -0.892674267, 0.370076209, -0.258745193, -0.598801553, -0.140952155, 0.697696567, ...]	[-0.02379784695825423, 0.11023467142025102, -0.0007805552248217122, 0.011769339381535295, 0.06963300920272536, -0.02548049511877664, -0.03757601190455139, 0.02308349306912037б, 0.02326446100630121, 0.04334829698291676, 0.01858458138884663, 0.050494730364757094, -0.03459691506409323, 0.06665788576904598, 0.01958634878280504б, -0.010353054505289684, 0.002184927380028947, 0.03553502232542259, -0.04129151717934945, 0.018676009432135442, 0.01642033251488061, -0.02721409152662376, 0.03884527958066426, -0.015657663571874134, -0.03757542108342231, -0.047251034012265505, 0.09308277990930837, -0.015896945051669466, 0.007921473652379598, -0.014876247892399786, 0.002252505637255518, 0.024681805553464303, 0.05313474615976798, 0.00395593962799294, 0.025742009574779425, 0.08715047209540527, 0.029629135996496143, 0.013657140858314142, -0.038415455593024665, -0.018295206933103544, -0.02438686513169107, -0.05229954596773251, 0.021203168227212115, 0.0020848080706018235, 0.034073821933961294, 0.0023610595845454767, 0.03481079522203354, -0.005773573319840027, 0.018461193765505205, 0.013062421818253558, 0.0431272839079299, 0.0545780456766709, 0.012173548022814803, 0.02611223881790076, -0.00016556535856625964, 0.009428528300202538, -0.023885173302747423, 0.050027284539761366, 0.040754829991318846, -0.030542003594687922, -0.052186517396215694, -0.029243390461626367, -0.07033108034944217, 0.012863155665836467, -0.05442734390766332, -0.035576818771987005, -0.02091300171199427, 0.04900606368669939, 0.006584174126444206, -0.04669635174154146, -0.051702262081564296, -0.016093796415770866, 0.00511570527028036, 0.020614616749164188, 0.026410300499212124, 0.032551988497190196, -0.001295900894142406, 0.01996361684258641, -0.011283138200188954, -0.028308948194854796, -0.030614357345964953, 0.03260538145157811, -0.017836990955294226, -0.005308543252896794, 0.026581119027299673, 0.005381423140266226, 0.03694070822977415, 0.08000826892101685, -0.023085278462331743, -0.04281865331656551, -0.01908915897494887, -0.04540884475272435, 0.0007693771871636435, -0.052404383854842716, -0.03206144792321064, 0.013291723018238092, -0.009293138423974067, -0.021506663200191876, -0.005062462696929981, 0.025058594132268568, ...]

Figure 5. The Results of Embedding and Normalization using IndoBERT

2.5 Data Processing

Clustering in this study was conducted using the Agglomerative Clustering Algorithm to explore the semantic structure of news documents represented by IndoBERT embeddings [19], [20]. Each document was encoded as a high-dimensional vector, and clustering was performed using various distance metrics to analyze semantic proximity.

Agglomerative Clustering was selected for its hierarchical, bottom-up approach, which does not require a predefined number of clusters—an advantage for large, thematically diverse corpora [21], [22]. Unlike K-Means, which depends on random initialization, this method builds a dendrogram that enables multi-level semantic interpretation and visualization of inter-document relationships.

This approach is well-suited for dense embeddings produced by transformer models like IndoBERT, where data often form complex, non-spherical manifolds [23], [24]. By iteratively merging the most similar document pairs, it identifies natural hierarchical groupings within semantic geometry. Empirically, hierarchical clustering has achieved higher topic coherence (0.68) compared to K-Means (0.55) and DBSCAN (0.59). Two distance metrics were applied: Cosine distance, capturing angular similarity, and Euclidean distance, measuring absolute geometric separation—both reflecting different aspects of semantic relationships within IndoBERT's embedding space:

$$d_{cosine}(x,y) = 1 - \frac{x \cdot y}{\|x\|\|y\|} \tag{1}$$

where $x.y = \sum_{i=1}^{n} x_i y_i$ represents the dot product, and $\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$ is the L2-norm of vector x.
The value of d_{cosine} ranges between 0 and 2, with 0 indicating perfect similarity (identical direction) and values approaching 2 indicating complete dissimilarity. The Cosine distance is defined as:

$$d_{cosine}(x,y) = 1 - \frac{x \cdot y}{\|x\|\|y\|} \tag{2}$$

Meanwhile, the Euclidean distance ($d_{euclidean}$) measures the straight-line distance between two points in the vector space and is defined as:
$d_{euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}^2$
This metric captures absolute geometric differences between embedding vectors, making it sensitive to magnitude as well as direction.

In hierarchical clustering, the distance between clusters is determined based on a linkage criterion (L(A,B)), which defines how distances among documents in different clusters are aggregated. The linkage function can be mathematically expressed as:

- Single linkage (nearest neighbor):

  $$L_{single}(A,B) = \frac{min}{x \in A, y \in B} d(x,y) \tag{3}$$

  This criterion tends to produce elongated, chain-like clusters.
- Average linkage (mean distance):

  $$L_{average}(A,B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x,y) \tag{4}$$

  This approach balances the tendencies of single and complete linkage by averaging all pairwise distances between members of two clusters.

In this study, the average linkage criterion was primarily adopted, as it provides a balanced trade-off between compactness and cluster separability, particularly effective in representing semantic relationships within dense IndoBERT embeddings.

After the distance between documents is calculated, cluster merging is performed iteratively according to the linkage (the rule for calculating the distance between clusters) used. In this research, cosine distance utilized the average linkage (the mean distance between all pairs of points across the two clusters), while euclidean distance employed the ward linkage (the increase in total variance when two clusters are merged). The pseudo-formula for Agglomerative Clustering is presented as follows:

```
$Start: C = \{\{X_1\}, \dots, \{X_n\}\} Distance\ Calculation\ D(C_i, C_j)\ for\ all\ i, j$
$While\ |C| > k:$
$(C_p, C_q) = \arg\min D(C_i, C_j)\ C \leftarrow (C \setminus \{C_p, C_q\}) \cup \{C_p \cup C_q\} Update\ D(C, \cdot)\ according\ to\ linkage$
```

In this study, several variables and notations are used to define the process of Agglomerative Clustering. The variable X_1 represents the $i-th$ document embedding vector generated by IndoBERT, where $i = 1,2,\dots,n$. Each vector $X_1 \in \mathbb{R}^{768}$ encodes the semantic representation of a news document. The set C denotes the collection of clusters at a given iteration, which initially consists of singleton clusters such that $C = \{\{X_1\}, \{X_2\}, \dots, \{X_n\}\}$. The notation $|C|$ refers to the number of clusters present at a particular stage, and the merging process continues until the cluster count equals the predetermined number k.

The distance between two clusters, $D(C_i, C_j)$, is computed based on the selected metric (Cosine or Euclidean) and the linkage criterion (single, complete, or average). The pair of clusters with the smallest distance value, identified as $(C_p, C_q) = \arg min\, D\, (C_i, C_j)$, represents the most similar clusters and will be merged during that iteration. The union of these clusters, $C_p \cup C_q$, orms a new cluster that replaces both within the set C. After each merge, the distance matrix $D(C_i.)$ is updated to recalculate distances between the new cluster and the remaining clusters according to the chosen linkage rule. Finally, the parameter k defines the desired number of clusters, determined empirically or through evaluation metrics such as the silhouette score or topic coherence.

3. RESULT AND ANALYSIS

3.1 Dimensionality Reduction Results (Semantic Geometry)

The Semantic Geometry of 35798 news documents is displayed in 2 dimensions using UMAP Projection as follows:

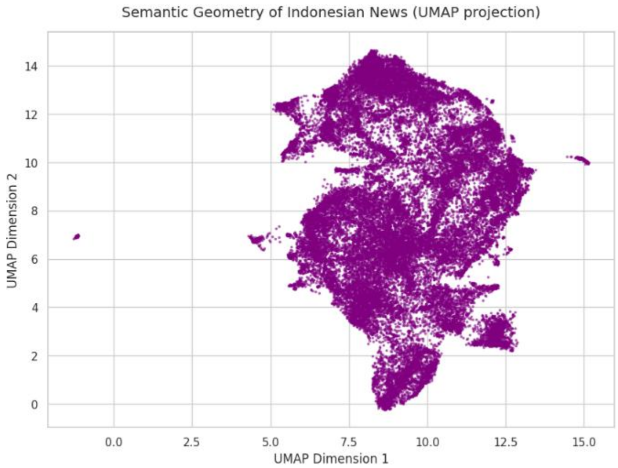


Figure 6. The Semantic Geometry of Indonesian News

From the semantic geometry results above, it is known that each point on the UMAP plot represents the meaning of a single news document. The distance between points illustrates the similarity of meaning; thus, semantically similar texts cluster closely, while dissimilar texts are separated widely. The overall shape, resembling an island, depicts the data distribution. Upon closer inspection, there is a large cluster of points in the center, which signifies a collection of news with common vocabulary (such as national politics, government, economy, etc.) and smaller clusters above and below, which represent distinct subtopics.

3.2 Optimal Clustering

In the initial stage, the evaluation for the optimal number of clusters was performed using the Silhouette Score based on Cosine Similarity, which indicated cluster=2 as the best result, yielding a Silhouette score of 0.8927.

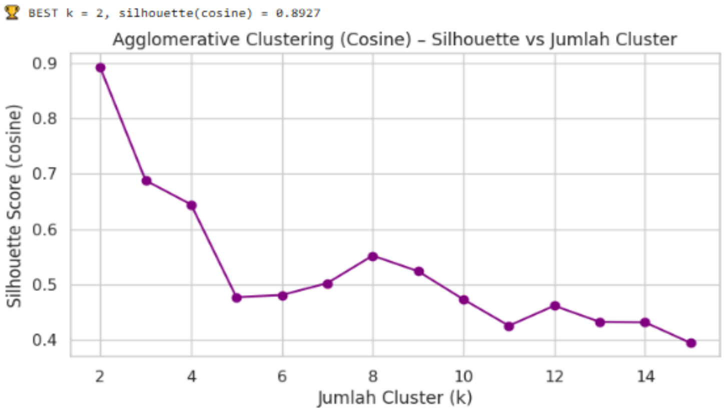

```
🏆 BEST k = 2, silhouette(cosine) = 0.8927
```

Figure 7. The Optimal Cluster by Cosine Silhouette Score

However, this clustering is not thematically representative enough because these two clusters are too broad to capture the finer variations in meaning within a news corpus containing over 35000 document data points. This is due to the density and homogeneity of the IndoBERT embeddings, which causes the similarity between vectors to be quite high. To obtain more granular clustering, the Elbow Method was employed, which resulted in cluster=6 as the optimal cluster number.

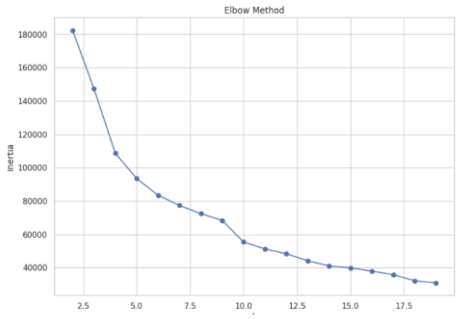


Figure 8. The Optimal Cluster by Elbow Method

The Elbow method works by finding the point on the inertia/WCSS (Within-Cluster Sum of Squares) plot against the number of clusters. From Figure 8, it can be observed that starting from cluster=6, the decrease in WCSS is no longer significant. This approach suggests that even though documents have high similarity within the semantic geometry (embedding space), additional grouping is still necessary to make the cluster interpretation more thematically representative.

3.3 Agglomerative Clustering Results

The first round of clustering was performed using the cosine metric within Agglomerative Clustering, with IndoBERT embeddings as the input. The resulting clustering with cluster=6 is visualized using UMAP below.

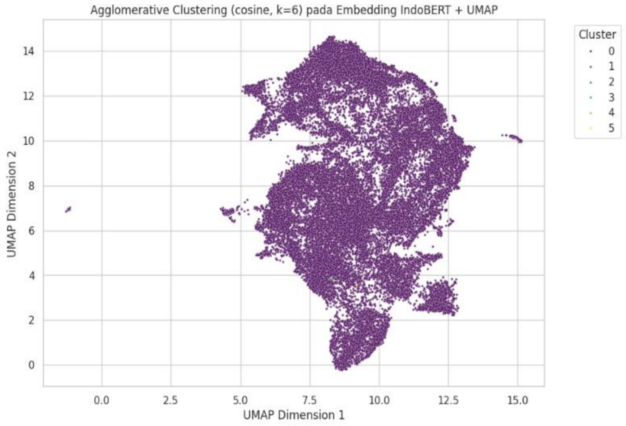


Figure 9. The Clustering Results with Cosine Metric

From the results above, the distribution of the news documents formed is as follows: cluster_0 = 35766 documents, cluster_1 = 21 documents, cluster_2 = 3 documents, cluster_3 = 2 documents, cluster_4 = 4 documents, and cluster_5 = 2 documents. The clustering result with cluster=6 based on cosine similarity shows a highly uneven distribution. Cluster_0 dominates the partition with 35766 documents, while the other clusters only contain 2 to 21 documents. This indicates that most news articles have very high semantic similarity, while the smaller clusters represent unique or outlier documents. The UMAP visualization reinforces this finding, as almost the entire plot appears in the color of cluster_0.

Next, clustering was performed using the euclidean metric within Agglomerative Clustering, also with IndoBERT embeddings as the input. The resulting clustering with cluster=6 is visualized using UMAP below.

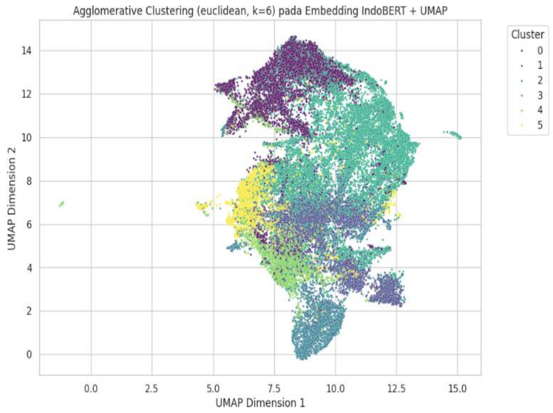


Figure 10. The Clustering Results with Euclidean Metric

From the results above, the distribution of the news documents formed is as follows: cluster_0 = 7851 documents, cluster_1 = 5822 documents, cluster_2 = 3784 documents, cluster_3 = 12815 documents, cluster_4 = 2566 documents, and cluster_5 = 2960 documents. The clustering result using the euclidean metric with cluster=6 shows a more even distribution of documents compared to clustering with the cosine metric. The clustering results also indicate that the use of the euclidean metric is capable of distinguishing documents into several major (cluster_3) and minor (cluster_4) clusters, allowing for better granularity in topic identification.

Subsequently, clustering was performed using the euclidean metric with Agglomerative Clustering, using inputs reduced by PCA and UMAP. The clustering results for each are presented as follows.

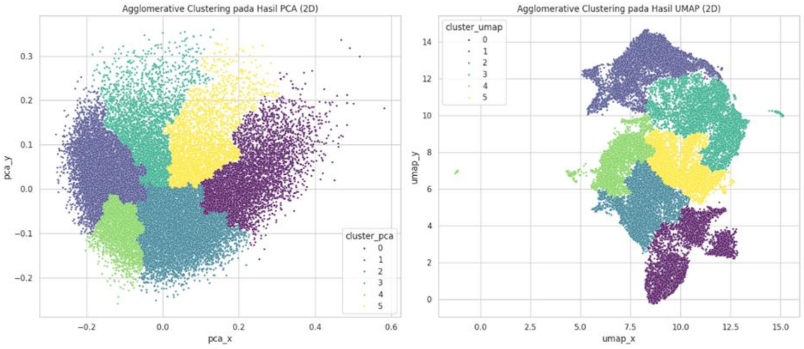


Figure 11. PCA and UMAP Clusterings Results with Euclidean Metric

For clustering with PCA input, the distribution of news documents formed is: cluster_0 = 5252 documents, cluster_1 = 8921 documents, cluster_2 = 10725 documents, cluster_3 = 3731 documents, cluster_4 = 3916 documents, and cluster_5 = 3253 documents. Meanwhile, the distribution of news documents with UMAP input is: cluster_0 = 5889 documents, cluster_1 = 8204 documents, cluster_2 = 5654 documents, cluster_3 = 7284 documents, cluster_4 = 4254 documents, and cluster_5 = 4513 documents. Clustering using the euclidean metric with embeddings reduced via PCA still shows a relatively uneven cluster distribution (although much better compared to the cosine metric) with the largest cluster containing 10725 documents. Conversely, clustering with UMAP demonstrates a more balanced distribution of documents among the clusters, with the number of documents per cluster ranging from 4254 to 8204 documents. This difference suggests that dimensionality reduction through UMAP can clarify the structure of minor clusters, making the identification of previously obscure topics easier.

3.4 Cluster Profiling

To further understand the characteristics of each cluster, profiling was conducted by analyzing the documents closest to the cluster centroid, as well as the dominant keywords based on TF-IDF (Term Frequency-Inverse Document Frequency) and n-gram analysis. Below is the cluster profiling based on the cosine metric with IndoBERT embeddings input, derived from the 10 data points closest to the centroid.

Table 1. Profiling of Cosine-Based Clustering with IndoBERT Embedding

Cluster	Size	Document Content	Top TF-IDF	Top Bigram	Top Trigram
0	35766	Economic news, government policies, development, support facilities for business actors, inflation, social assistance, and socio-economic activities.	"health", "which/that", "and", "in/at", "also", "for", "this", "with", "in/within", "from"	economic growth, in Indonesia	President Joko Widodo, COVID-19 Pandemic
1	21	Transportation, train and bus ticket prices, travel schedules, public services related to holidays.	'train', 'ticket', 'BPJS', 'price', 'Nataru', 'Jakarta', 'health', '2022', 'from', 'and'	train, ticket price, train ticket, Christmas holidays, Senen Market	train ticket prices, train tickets, for the Christmas holidays
2	3	Entertainment, public figures' profiles, gimmick terms, political branding, artists.	'style', 'public', 'Queen Sofya', 'junior high school window', 'Sofya', 'who', 'and'	Queen Sofya, junior high school window, entertainment world, blunt	This kind of style, peeking through the junior high school window, to attract attention.
3	2	Environment, air pollution, innovative solutions for cities, market evictions as a local/global phenomenon.	'pollution', 'air pollution', 'air', 'market', 'solution', 'in', 'and'	air pollution, entertainment center, one of	is the best solution, occurred in England, has studied pollution

4	4	Research and technological innovation (food, music), educational and community activities (breastfeeding, parenting).	'breast milk', ‘technology’, ‘violin’, ‘research’, ‘Nasir’, ‘mother’, ‘in’, 'and'	processing technology, technology research, Southeast Asia	food processing technology, Southeast Asia developed, industrial processing technology
5	2	Technology, vehicle safety, virtual fashion shows, green screen technology.	'virtual', ‘safety’, ‘headrest’, ‘screen’, ‘green screen’, ‘green technology’, ‘that’, ‘now’, 'with'	green screen, green technology, headrest, virtual clothing	green screen technology, virtual clothing, virtual fashion show

From the profiling results above, it can be concluded that the thematic labels are: cluster_0 is General News, Public Policy, and Economy; cluster_1 is Transportation and Public Services; cluster_2 is Entertainment and Public Figures; cluster_3 is Environmental Issues; cluster_4 is Research and Technology Innovation; and cluster_5 is Technology and Virtualization.

Next, the cluster profiling based on the euclidean metric with IndoBERT embeddings input, derived from the 10 data points closest to the centroid, is displayed.

Table 2. Profiling of Euclidean-Based Clustering with IndoBERT Embedding

Cluster	Size	Document Content	Top TF-IDF	Top Bigram	Top Trigram
0	7851	National political news, parties, elections, political figures	politics, party, Prabowo, who, and	political party, chairman, currently, money politics, 2024 election, Prabowo Subianto, politics	party chairman, President Joko Widodo, vice presidential candidate, Gibran Rakabuming Raka.
1	5822	Digital transformation, technology-based economy, industrial innovation	technology, Indonesia, that, and, for	in Indonesia, one of, currently, more, besides that, cooperation, this matter	in a press release, becoming one of the reasons why human resources, the press that is accepted, is one of the reasons why.
2	3784	Public health, lifestyle, nutrition, disease and prevention	health, children, who, and, for	for health, BPJS Health, mental health, more, one, which can, in addition, can help	blood sugar levels, therefore, to maintain health, lose weight.
3	12815	Education and health, public services, national social issues	health, education, school, which, and	one of them, in Indonesia, COVID-19, children, entertainment venues, among others	free health check, nightlife venues, therefore, the COVID-19 pandemic.
4	2566	Green economy, sustainable development, clean energy, companies	health, which, and, in, for	In Indonesia, one of the current factors contributing to economic growth is...	billion dollars, therefore, human resources, in a press release, became one of the accepted press releases.
5	2960	National economy, economic growth, fiscal policy, global risks	economy, growth, percent, Indonesia, which	economic growth, Indonesian economy, US dollar, percent, interest rate, currently, higher	Indonesia's economic growth, billion US dollars, household consumption, percent annually, benchmark interest rate, United States.

From the profiling results above, it can be concluded that the thematic labels are: cluster_0 is National Politics and Elections; cluster_1 is Digital Transformation and Industry; cluster_2 is Public Health and Lifestyle; cluster_3 is Public Services, Education, and Social Issues; cluster_4 is Green Economy and Sustainable Development; and cluster_5 is Macroeconomics and Fiscal Policy.

Subsequently, the cluster profiling based on the euclidean metric with PCA input, derived from the 10 data points closest to the centroid, is displayed.

Table 3. Profiling of Euclidean-Based Clustering with PCA

Cluster	Size	Document Content	Top TF-IDF	Top Bigram	Top Trigram
0	5252	Digital economy, Industry 4.0 technology, business and education transformation, startup innovation	which, and, in, for, also, health, with, this, in, from	one of them, more, for health, in Indonesia, besides that, this thing, which can, currently	Therefore, one of them is blood sugar level.
1	8921	National politics, parties, government, political figures, legislative issues	who, in, and, politics, also, that, party, this, for, not	political party, currently, chairman,	President Joko Widodo, party chairman, nightlife

				one of, money politics, none, he said	venue, Jakarta politics Kompas
2	10725	National economy, MSMEs, education and human resource transformation, sustainable development, public health	and, who, in, for, this, also, with, health, Indonesia, in	in Indonesia, one of, economic growth, currently, more, in addition, this	billion dollars, becoming one of the reasons why, therefore, in a press release, human resources, the COVID-19 pandemic
3	3731	Politics, elections, campaigns, political parties and figures, the dynamics of identity politics	who, and, in, politics, also, party, that, this, for, with	chairman, political party, currently, one of, none, Prabowo Subianto, in Indonesia	party chairman, President Joko Widodo, vice presidential candidate, therefore
4	3916	MBG program (free nutritious meals), economic stimulus, public health, sustainable development	and, which, in, percent, economy, also, this, health, for, Indonesia	economic growth, currently, COVID-19, in Indonesia, one of, free healthcare, in addition	free health checkups, Indonesia's economic growth, the COVID-19 pandemic, and the creative economy
5	3253	Politics, education and human resources, the role of young people, development and reform agenda	who, and, in, also, for, this, with, in, from, child	Currently, one of the issues in Indonesia is that children are not only...	Therefore, those who are in, become one of, on the other hand, ministers, President Joko Widodo

From the profiling results above, it can be concluded that the thematic labels are: cluster_0 is Technology, Business Transformation, and Innovation; cluster_1 is National Politics and Government; cluster_2 is National Economy, Human Resources, and Development; cluster_3 is Electoral Politics and Campaigning; cluster_4 is Economy, Social Programs, and Public Health; and cluster_5 is Politics, Education, and Development Agenda.

Finally, the cluster profiling based on the euclidean metric with UMAP input, derived from the 10 data points closest to the centroid, is displayed.

Table 4. Profiling of Euclidean-Based Clustering with UMAP

Cluster	Size	Document Content	Top TF-IDF	Top Bigram	Top Trigram
0	5889	Health and Technology, AI and medical technology, lifestyle, dengue fever, technology adaptation, hybrid working, and smart cars	which, and, in, for, also, with, this, health, technology, can	more, one, for health, mental health, besides that, which can, which is able to, in Indonesia	blood sugar levels, to maintain health, read also whether, what is in, becomes one of
1	8204	Political dynamics, political promises, ministers' political ethics, political party coalitions, inquiry rights, discourse on postponing elections, and parliamentary thresholds	who, and, in, politics, party, also, that, in, this, for	chairman, money politics, currently, 2024 elections, politics, Prabowo Subianto	party chairman, President Joko Widodo, Jakarta politics, Kompas, vice presidential candidate, Gibran Rakabuming Raka
2	5654	Digital economy value, recession mitigation, green economy, banking sector, risk management industry, MSMEs, advertising technology, and environmentally friendly development.	and, which, in, for, also, Indonesia, this, with, economy, within	in Indonesia, one of the current economic growths, cooperation, more than that	billion dollars, in a press release, therefore, becomes one of the human resources, press received, small and medium
3	7284	Education, social issues, health disorders, protection of parties in education, violations of PPKM/entertainment venues, KPAI, certificate whitening, SPMB, children not attending school, and military education	at, who, and, school, education, health, also, this, for, children	entertainment venues, Jakarta Kompas, COVID-19, children, currently, public schools, one of, education office	nightlife venues, free health checks, he also reads, education and culture, located in, the COVID-19 pandemic, head of the education office
4	4254	Macroeconomics and finance, threat of crisis, Indonesia's economic	which, and, economy, in, percent,	economic growth, Indonesian economy, currently, BPJS	Indonesia's economic growth, billion US dollars, household

		growth, draft state budget, global recession, inflation, Sri Mulyani, and ADB	Indonesia, growth, also, this, on	Health, US dollar, percent, interest rate	consumption, percent annually, percent read also
5	4513	Vocational education transformation, students/agents of national transformation, virtual edu expo, computer science, BPJS Health, and social assistance distribution	and, who, in, health, for, this, also, education, with, in	In Indonesia, one of them, currently, BPJS Kesehatan, cooperation, in addition to that,	free health checks, tourism and economy, creative economy, human resources, in a press release, therefore, the press received

From the profiling results above, it can be concluded that the thematic labels are: cluster_0 is Health and Technology; cluster_1 is Politics; cluster_2 is Economy and Industry; cluster_3 is Education and Social Issues; cluster_4 is Macroeconomics and Finance; and cluster_5 is Vocational Education Transformation and Social Security.

3.5 Discussion

The findings of this study indicate that the choice of distance metrics and dimensionality reduction techniques plays a decisive role in shaping the granularity and thematic interpretability of Indonesian online news clusters derived from IndoBERT embeddings. Empirical results show that the use of Cosine similarity within Agglomerative Clustering leads to a highly skewed cluster distribution, where more than 99% of documents are concentrated in a single dominant cluster. Although this configuration achieved a relatively favorable Silhouette Score, it proved ineffective in capturing meaningful subtopic variation, as the remaining clusters merely represented marginal or outlier themes such as transportation and entertainment.

In contrast, applying Euclidean distance produced a noticeably more balanced clustering structure. This suggests that Euclidean distance is better suited to distinguishing documents that share similar semantic orientations but differ in contextual emphasis or intensity. When combined with PCA, cluster balance improved slightly; however, PCA's linear projection was still limited in preserving the complex, non-linear semantic relationships encoded in transformer-based embeddings such as IndoBERT [25], [26].

The most substantial improvement was observed when Euclidean distance was combined with UMAP [27], [28]. Rather than emphasizing UMAP's mathematical formulation, this study highlights its empirical contribution: UMAP effectively redistributes dense, high-dimensional embeddings into a lower-dimensional space while maintaining local semantic relationships. As a result, the Euclidean–UMAP configuration achieved the highest topic coherence score (0.71) along with a well-balanced cluster size distribution, enabling clearer thematic separation and more intuitive interpretation of clusters.

Empirically, this configuration revealed five coherent thematic domains: Politics; Health and Technology; Macroeconomics and Finance; Economy and Industry; and Education and Social Issues [29], [30]. The relatively even distribution across these themes demonstrates that non-linear dimensionality reduction is essential for uncovering latent thematic boundaries in large-scale Indonesian news corpora. Overall, the results confirm that combining Euclidean distance with UMAP offers a practical and empirically robust approach for enhancing thematic clarity and balance in semantic clustering, without the need for extensive theoretical complexity.

4. CONCLUSION

This study demonstrated that the choice of clustering metrics and dimensionality reduction methods significantly affects the thematic organization of Indonesian news using IndoBERT embeddings. The Cosine similarity metric produced a skewed result with over 99% (35,766) documents in one cluster (Silhouette Score = 0.61), failing to capture topic diversity. In contrast, the Euclidean + UMAP configuration achieved the most balanced distribution (4,254–8,204 documents per cluster) and the highest topic coherence (0.71), showing better separation of dense semantic subtopics while maintaining contextual relations.

The UMAP–Euclidean model identified five main themes Politics, Health & Technology, Macroeconomics & Finance, Economy & Industry, and Education & Social Issues reflecting Indonesia's current discourse landscape. For future work, the study suggests integrating GPU-accelerated UMAP or approximate nearest-neighbor search (ANN) for faster processing of large corpora and exploring multimodal embeddings to enrich topic granularity. Computationally, the model shows moderate scalability (average 0.42 s/document on a 12-core CPU, 16 GB GPU), yet larger applications require distributed or adaptive frameworks to maintain efficiency and accuracy. In summary, the Euclidean + UMAP configuration provides the most effective, balanced, and reproducible approach for semantic clustering of Indonesian text corpora.

5. REFERENCES

[1] D. Gandasari and D. Dwidienawati, “Content analysis of social and economic issues in Indonesia during the COVID-19 pandemic,” *Heliyon*, vol. 6, no. 11, 2020.

[2] E. H. Susanto, R. Loisa, and A. Junaidi, “Cyber media news coverage on diversity issues in Indonesia,” *J. Hum. Behav. Soc. Environ.*, vol. 30, no. 4, pp. 510–524, 2020.

[3] D. H. Santoso, “New media and nationalism in Indonesia: An analysis of discursive nationalism in online news and social media after the 2019 Indonesian presidential election,” *J. Komun. Malaysian J. Commun.*, vol. 37, no. 2, pp. 289–304, 2021.

[4] T. Sommerschield *et al.*, “Machine learning for ancient languages: A survey,” *Comput. Linguist.*, vol. 49, no. 3, pp. 703–747, 2023.

[5] D. G. I. Purnawati, D. P. S. Putri, and I. N. Piarsa, “Implementation of Text Mining for Evaluating the Relevance Between News Headlines and Content on a Web-Based Platform,” *J. Appl. Informatics Comput.*, vol. 9, no. 4, pp. 1463–1476, 2025.

[6] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, and D. F. Wong, “A survey on llm-generated text detection: Necessity, methods, and future directions,” *Comput. Linguist.*, vol. 51, no. 1, pp. 275–338, 2025.

[7] D. U. K. Putri and D. N. Pratomo, “Clickbait detection of Indonesian news headlines using fine-tune bidirectional encoder representations from transformers (BERT),” *Inf. J. Ilm. Bid. Teknol. Inf. Dan Komun.*, vol. 7, no. 2, pp. 162–168, 2022.

[8] D. Pawar, S. Phansalkar, A. Sharma, G. K. Sahu, C. K. Ang, and W. H. Lim, “Survey on the biomedical text summarization techniques with an emphasis on databases, techniques, semantic approaches, classification techniques, and similarity measures,” *Sustainability*, vol. 15, no. 5, p. 4216, 2023.

[9] S. Medileh *et al.*, “Optimizing deep learning for webshell detection based on flexible dataset reduction,” *Egypt. Informatics J.*, vol. 31, p. 100770, 2025.

[10] A. Suryadibrata and J. C. Young, “Embedding from Language Models (ELMos)-based Dependency Parser for Indonesian Language.,” *Int. J. Adv. Soft Comput. Its Appl.*, vol. 13, no. 3, 2021.

[11] K. W. Trisna, J. Huang, H. Liang, and E. M. Dharma, “Fusion text representations to enhance contextual meaning in sentiment classification,” *Appl. Sci.*, vol. 14, no. 22, p. 10420, 2024.

[12] M. Sitopu, W., Nababan, E., & Budiman, “Reducing Semantic Distortion of Multiword Expressions for Topic Modeling with Latent Dirichlet Allocation.,” *J. Inf. Syst. Informatics*, vol. 7, no. 3, pp. 2920–2938, 2025, doi: https://doi.org/10.51519/journalisi.v7i3.1266.

[13] U. Orhan and C. N. Tulu, “A novel embedding approach to learn word vectors by weighting semantic relations: SemSpace,” *Expert Syst. Appl.*, vol. 180, p. 115146, 2021.

[14] Y. Yin, Y. Zhang, Z. Liu, S. Wang, R. R. Shah, and R. Zimmermann, “GPS2Vec: Pre-trained semantic embeddings for worldwide GPS coordinates,” *IEEE Trans. Multimed.*, vol. 24, pp. 890–903, 2021.

[15] P. Poschmann, J. Goldenstein, S. Büchel, and U. Hahn, “A vector space approach for measuring relationality and multidimensionality of meaning in large text collections,” *Organ. Res. Methods*, vol. 27, no. 4, pp. 650–680, 2024.

[16] C. Moreno Pérez and M. Minozzo, “Natural language processing and financial markets: semi-supervised modelling of coronavirus and economic news,” 2022.

[17] T. A. Mohd Tajul Ariffin, S. N. H. Sheikh Abdullah, F. Fauzi, Z. Murah, and M. K. Hasan, “Review on honeynet analysis: can LSTM and shot learning drive intelligent cyber threat modelling and automation?,” *Cluster Comput.*, vol. 28, no. 9, pp. 1–32, 2025.

[18] Y. Du, H. Sun, and M. Abdollahi, “Toward deep multi-view document clustering using enhanced semantic embedding and consistent context semantics,” *Knowl. Inf. Syst.*, vol. 67, no. 2, pp. 1073–1100, 2025.

[19] S. Vahidnia, A. Abbasi, and H. A. Abbass, “Embedding-based Detection and Extraction of Research Topics from Academic Documents Using Deep Clustering.,” *J. Data Inf. Sci.*, vol. 6, no. 3, pp. 99–122, 2021.

[20] M. Q. Memon, Y. Lu, P. Chen, A. Memon, M. S. Pathan, and Z. A. Zardari, “An ensemble clustering approach for topic discovery using implicit text segmentation,” *J. Inf. Sci.*, vol. 47, no. 4, pp. 431–457, 2021.

[21] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, and N. Nobani, “Embeddings evaluation using a novel measure of semantic similarity,” *Cognit. Comput.*, vol. 14, no. 2, pp. 749–763, 2022.

[22] J. Gómez and P.-P. Vázquez, “An empirical evaluation of document embeddings and similarity metrics for scientific articles,” *Appl. Sci.*, vol. 12, no. 11, p. 5664, 2022.

[23] V. Gupta, A. Dixit, and S. Sethi, “A comparative analysis of sentence embedding techniques for document ranking,” *J. Web Eng.*, vol. 21, no. 7, pp. 2149–2185, 2022.

[24] S. Khosla, S. Jain, M. A. Anupama, and R. S. R. Thavva, “Comparative Analysis of Multiple Embedding Models for Text Based Document Similarity,” in *International Conference on Artificial Intelligence and Speech Technology*, Springer, 2024, pp. 169–180.

[25] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, “A survey of text representation and embedding techniques in nlp,” *IEEe Access*, vol. 11, pp. 36120–36146, 2023.

[26] M. Boyapati and R. Aygun, “Semanformer: Semantics-aware Embedding Dimensionality Reduction Using Transformer-Based Models,” in *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, IEEE, 2024, pp. 134–141.

[27] E. Becht *et al.*, “Dimensionality reduction for visualizing single-cell data using UMAP,” *Nat. Biotechnol.*, vol. 37, no. 1, pp. 38–44, 2019.

[28] Y. Hozumi, R. Wang, C. Yin, and G.-W. Wei, “UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets,” *Comput. Biol. Med.*, vol. 131, p. 104264, 2021.

[29] X. He and H. Zhao, “Macro-level insights into the digital economy: topic identification and trend analysis,” *Appl. Econ.*, pp. 1–18, 2025.

[30] L. LinLin, “Examining Macroeconomic Policies And Their Impact On Corporate Finance: A Comparative Assessment Of Developed And Emerging Economies,” *Int. J. Econ. Financ. Stud.*, vol. 16, no. 3, pp. 365–388, 2024.