JURNAL LOCUS: Penelitian & Pengabdian
Volume 4, No. 10, Oktober 2025
E-ISSN 2829-7334| P-ISSN 2829-5439
Hompage: https://locus.rivierapublishing.id/index.php/jl

Klasterisasi Judul Berita Online Isu Pemilu Prabowo Subianto dengan Kombinasi LLMS Embedding Dengan HDBSCAN

Nanda Perdana*, Handri Santoso
Universitas Pradita, Indonesia
Email: nanda.perdana@student.pradita.ac.id*, handri.santoso@pradita.ac.id

Abstrak:
Penelitian ini bertujuan untuk mengelompokkan judul-judul berita politik daring yang berkaitan dengan Presiden Prabowo Subianto selama Pemilu 2024 menggunakan pendekatan berbasis embedding large language models (LLMS) dan algoritma klasterisasi HDBScan. Data yang digunakan dalam penelitian ini berjumlah 24.000 judul berita yang kemudian dianalisis mellaui beberapa tahap meliputi pra-pemrosesan teks, ekstraksi embedding menggunakan model OpenAI, reduksi dimensi menggunakan UMAP, serta klasterisasi berbasis densitas adaptif dengan HDBSCAN. Hasil penelitian menunjukkan terbentuknya 85 klaster tematik dan identifikasi sekitar 27,2% data sebagai noise. Hasil temuan pada penelitian ini mengindikasikan bahwa kombinasi embedding LLM dan HDBSCAN efektif dalam, mengungkap struktur semantik wacana politik digital dari data yang digunakan, serta mampu menangani karakteristik data teks pendek yang kompleks dan heterogen. Pendekatan ini memberikan kontribusi metodologis terhadap studi analisis media berbasis data besar dan menwarakan landasan bagi penelitian lanjutan dalam pemetaan itu publik di ruang digital. Hasil penelitian ini dapat digunakan sebagai sarana untuk penelitian lebih lanjut dengan studi kasus yang berbeda namun menggunakan algoritma yang sama.

Kata kunci: *Analisis Wacana Digital; HDBSCAN; Klasterisasi Teks Pendek; LLM; Pemilu 2024; Prabowo Subianto*

Abstract
This study aims to cluster online political news headlines related to President Prabowo Subianto during the 2024 General Election using an embedding-based approach with Large Language Models (LLMs) and the HDBSCAN clustering algorithm. The dataset consists of 24,000 news headlines, which were analyzed through several stages including text preprocessing, embedding extraction using an OpenAI model, dimensionality reduction with UMAP, and adaptive density-based clustering with HDBSCAN. The results of the study produced 85 thematic clusters and identified approximately 27.2% of the data as noise. These findings indicate that the combination of LLM embeddings and HDBSCAN is effective in uncovering the semantic structure of digital political discourse from the dataset while being capable of handling the complex and heterogeneous characteristics of short-text data. This approach provides a methodological contribution to big data-based media analysis studies and offers a foundation for future research in mapping public discourse within digital spaces. The findings of this study can be applied in further research with different case studies while employing the same algorithmic framework.

Keywords: *Digital Discourse Analysis; HDBSCAN; Short Text Clustering; LLM; 2024 General Election; Prabowo Subianto*

Corresponding: Nanda Perdana
E-mail: nanda.perdana@student.pradita.ac.id

PENDAHULUAN

Perkembangan teknologi pemrosesan bahasa alami atau Natural Language Processing (NLP), khususnya melalui pemanfaatan Large Language Models (LLMs) seperti model dari OpenAI, telah membuka peluang baru dalam analisis wacana politik digital. Dalam konteks pemilihan umum (Pemilu) 2024 di Indonesia, diskursus publik yang tersebar dalam berbagai kanal media daring, termasuk melalui judul-judul berita politik, menjadi sumber penting untuk memahami konstruksi opini publik dan dinamika isu yang berkembang. Pemanfaatan embedding berbasis LLM memungkinkan representasi semantik teks dalam bentuk vektor berdimensi tinggi yang dapat digunakan untuk mengukur kedekatan makna antar teks (Motoki

et al., 2023). Sebagaimana dikemukakan oleh Korade et al. (2025), model embedding dari OpenAI mampu menangkap nuansa kontekstual yang tersembunyi dalam teks pendek seperti judul berita, sehingga sangat relevan untuk kebutuhan klasterisasi tematik.

Salah satu pendekatan klasterisasi yang sesuai dengan karakteristik data teks pendek dan dinamis seperti berita politik adalah Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). Algoritma ini memiliki keunggulan dalam menangani data dengan kepadatan yang tidak seragam, sebuah karakteristik umum dalam korpus berita politik yang sering kali memuat topik-topik dominan berdampingan dengan isu-isu pinggiran. Penelitian sebelumnya (Chajia & Nfaoui, 2024; Hidayat, 2024) menunjukkan bahwa HDBSCAN efektif dalam mengidentifikasi kelompok narasi seperti isu kecurangan pemilu, politik dinasti, hingga gerakan protes, yang muncul sebagai tema penting dalam pemberitaan seputar Pemilu 2024 di Indonesia.

Metodologi klasterisasi yang dikembangkan dalam studi ini mencakup serangkaian tahap yang meliputi pengumpulan data berita secara sistematis, pra-pemrosesan teks untuk normalisasi data, serta ekstraksi fitur semantik menggunakan embedding berbasis LLM. Selanjutnya, untuk meningkatkan efisiensi komputasi dan menjaga struktur semantik data, teknik reduksi dimensi seperti Uniform Manifold Approximation and Projection (UMAP) diterapkan sebelum dilakukan klasterisasi dengan HDBSCAN. Strategi ini didukung oleh berbagai studi (Blanco-Portals et al., 2021), yang menekankan pentingnya reduksi dimensi dalam menangani embedding berdimensi tinggi yang berasal dari model bahasa besar.

Hasil dari proses klasterisasi ini memberikan wawasan tematik mengenai lanskap isu politik elektoral di Indonesia, dengan mengungkap pola-pola narasi yang tersebar dalam media. Analisis lanjutan terhadap klaster yang terbentuk dapat digunakan untuk memahami bagaimana persepsi publik terbentuk melalui media, serta mengidentifikasi isu-isu utama yang berpotensi mempengaruhi sikap pemilih. Dengan demikian, metode ini tidak hanya berkontribusi dalam konteks penelitian media dan politik digital, tetapi juga memberikan pendekatan berbasis data besar (big data) yang dapat digunakan untuk pengawasan dinamika wacana publik secara berkelanjutan. Secara keseluruhan, integrasi antara model LLM dan algoritma klasterisasi HDBSCAN menawarkan pendekatan yang komprehensif dan adaptif untuk menganalisis teks pendek dalam domain politik. Kombinasi ini memungkinkan eksplorasi yang lebih dalam terhadap karakter semantik teks berita serta mendukung deteksi isu-isu krusial dalam ekosistem informasi politik yang cepat berubah.

Meskipun berbagai studi telah dilakukan untuk menganalisis wacana politik di media digital, sebagian besar pendekatan yang digunakan masih berfokus pada metode topic modeling konvensional seperti Latent Dirichlet Allocation (LDA) atau teknik klasifikasi berbasis supervised learning. Metode-metode tersebut memiliki keterbatasan dalam menangani teks pendek, seperti judul berita, yang sering kali tidak mengandung cukup konteks untuk dianalisis secara efektif dengan pendekatan statistik tradisional. Selain itu, banyak penelitian yang masih mengandalkan teknik representasi teks berbasis frekuensi kata (seperti TF-IDF), yang cenderung gagal menangkap kedalaman makna semantik dan konteks diskursif dari wacana politik yang kompleks.

Di sisi lain, perkembangan teknologi LLM (Large Language Models) seperti OpenAI telah membuka peluang untuk membangun representasi semantik teks yang lebih canggih. Namun, pemanfaatan embedding LLM dalam konteks klasterisasi isu politik, khususnya di Indonesia dan dalam format teks pendek seperti judul berita, masih sangat terbatas. Terlebih lagi, hanya sedikit penelitian yang mengkombinasikan embedding LLM dengan algoritma klasterisasi yang adaptif seperti HDBSCAN dalam menganalisis narasi politik elektoral secara unsupervised. Hal ini menciptakan ruang penelitian yang signifikan untuk mengeksplorasi efektivitas pendekatan semantik-modern dalam memahami lanskap diskursif publik selama

Pemilu 2024, khususnya yang berpusat pada figur Presiden Prabowo. Berkaitan dengan hal tersebut, beberapa pertanyaan muncul dalam penelitian ini. Pertama, bagaimana cara mengelompokkan judul-judul berita politik terkait Presiden Prabowo selama Pemilu 2024 secara semantik. Kedua, seberapa efektif algoritma HDBSCAN dalam membentuk klaster tematik dari data teks pendek yang memiliki variasi makna dan kepadatan informasi yang tinggi. Ketiga, apa saja tema dominan atau isu utama yang muncul dalam klaster atau kelompok yang terbentuk. Pertanyaan-pertanyaan tersebut dapat berpengaruh pada kontribusi terhadap beberapa aspek meliputi metodologis, substantif, dan praktis. Pada kontribusi metodologis, penelitian ini menghadirkan pendekatan unsupervised semantic clustering dengan kombinasi embedding dari model bahasa besar dan dapat menjadi alternatif dari metode topik modeling konvensional. Pada kontribusi substantif, temuan pada penelitian ini dapat memperkaya literatur terkait analisis media dan politik digital di Indonesia. Pada kontribusi praktis, pendekatan yang dikembangkan dalam penelitian ini dapat digunakan oleh peneliti media, analis kebijakan, hingga lembaga pengawas pemilu untuk memahami pola penyebaran isu dan narasi politik digital.

Dua penelitian terdahulu memberikan pijakan penting, tetapi masih menyisakan ruang untuk eksplorasi lebih lanjut. Penelitian oleh Santoso dan Wibisono (2021) menekankan penggunaan metode topic modeling tradisional seperti Latent Dirichlet Allocation (LDA) dalam menganalisis wacana politik digital di Indonesia, namun metode ini terbatas dalam menangkap makna semantik mendalam dari teks pendek seperti judul berita, sehingga sering gagal mengungkap konteks diskursif yang kompleks. Sementara itu, studi oleh Firdaus et al. (2022) meneliti narasi politik di media sosial menggunakan teknik supervised learning, tetapi pendekatan ini sangat bergantung pada data berlabel dan tidak cukup adaptif dalam menangani dinamika isu politik yang terus berubah. Kelemahan dari dua penelitian ini adalah absennya pemanfaatan teknologi Large Language Models (LLMs) dan algoritma klasterisasi adaptif seperti HDBSCAN untuk analisis teks pendek yang lebih semantik.

Tujuan penelitian ini adalah mengidentifikasi tema dominan yang membentuk opini publik melalui media digital, sekaligus menguji efektivitas metode semantik-modern dalam konteks politik Indonesia. Manfaat praktis penelitian ini diharapkan dapat memberikan kontribusi bagi akademisi, analis kebijakan, dan lembaga pengawas pemilu dalam memahami pola penyebaran isu politik serta meningkatkan kualitas pengawasan dinamika wacana publik di era digital.

METODE PENELITIAN

Penelitian mengenai analisis berita berbasis unsupervised learning telah mengalami perkembangan pesat seiring dengan kemajuan teknologi pemrosesan bahasa alami (Natural Language Processing), khususnya melalui pemanfaatan Large Language Models (LLMs) (Lewis et al., 2020; Campello et al., 2015). Salah satu pendekatan yang mulai banyak diterapkan adalah penggabungan embedding berbasis LLM dengan algoritma klasterisasi adaptif seperti HDBSCAN. Embedding berbasis LLM, seperti yang dikembangkan oleh OpenAI, mampu merepresentasikan makna semantik teks secara mendalam (Sadeghi et al., 2024), bahkan untuk teks pendek seperti judul berita. Model ini memungkinkan representasi vektor dari teks yang tidak hanya menangkap informasi literal, tetapi juga nuansa kontekstual dan hubungan semantik yang tidak eksplisit. Sebagaimana dijelaskan oleh Keraghel et al. (2024), embedding dari LLM mampu mengidentifikasi kemiripan konseptual antar judul berita meskipun menggunakan kosakata yang berbeda, sehingga meningkatkan efektivitas proses klasterisasi. Temuan ini diperkuat oleh penelitian lain (Yang et al., 2024), yang menunjukkan bahwa pendekatan berbasis LLM memiliki performa tinggi dalam pengelompokan dokumen pendek dalam domain informasi dinamis seperti berita daring.

Untuk tahap klasterisasi, algoritma HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) banyak direkomendasikan karena kemampuannya dalam mengelompokkan data berdasarkan kepadatan tanpa memerlukan penentuan jumlah klaster secara eksplisit. Weng et al. (2022) menekankan bahwa HDBSCAN sangat sesuai untuk data teks yang jarang (sparse), seperti judul berita, karena kemampuannya dalam mendeteksi noise serta menghasilkan klaster dengan kepadatan yang konsisten. Keunggulan lain dari HDBSCAN adalah fleksibilitas parameter yang memungkinkan penyesuaian terhadap struktur data yang tidak seragam, suatu kondisi yang umum dalam korpus berita politik.

Beberapa studi juga menekankan pentingnya integrasi teknik reduksi dimensi sebelum proses klasterisasi untuk meningkatkan efisiensi dan interpretabilitas hasil. Blanco-Portals et al. (2021) menyarankan penggunaan UMAP (Uniform Manifold Approximation and Projection) untuk menyederhanakan representasi embedding berdimensi tinggi tanpa kehilangan struktur semantik yang penting. Dalam konteks ini, pipeline klasterisasi yang umum melibatkan empat tahap utama, yaitu: (1) pembersihan dan pra-pemrosesan data teks, (2) ekstraksi embedding menggunakan LLM, (3) reduksi dimensi dengan UMAP, dan (4) penerapan HDBSCAN untuk identifikasi klaster berbasis semantik.

Dengan mengintegrasikan embedding dari OpenAI dan klasterisasi HDBSCAN, pendekatan ini menawarkan solusi komprehensif dan adaptif untuk pengelompokan teks pendek dalam domain berita. Kombinasi tersebut terbukti mampu menangani tantangan variasi semantik, sparsity teks, serta kebutuhan skalabilitas dalam pengolahan data besar. Studi-studi yang dilakukan oleh Keraghel et al. (2024), Weng et al. (2022), Yang et al. (2024), serta ditunjang oleh penelitian lain secara kolektif mendukung efektivitas metode ini dalam membangun sistem klasterisasi yang robust, relevan, dan aplikatif dalam konteks analisis wacana politik digital.

Penelitian ini mengintegrasikan pendekatan teknologi natural language processing (NLP) dengan metode klasterisasi untuk mengidentifikasi struktur tematik dalam pemberitaan politik daring. Kerangka konsep penelitian ini dibangun atas dasar asumsi bahwa judul berita sebagai unit teks pendek memuat makna semantik yang dapat ditangkap dan dianalisis secara otomatis melalui representasi vektor.

Secara umum, kerangka penelitian ini terdiri dari lima komponen utama yang saling berkaitan. Pertama, objek analisis dalam studi ini adalah judul-judul berita daring yang berkaitan dengan Presiden Prabowo Subianto selama Pemilu 2024. Judul-judul tersebut dipandang sebagai representasi ringkas dari narasi publik dan isu-isu dominan yang beredar di media. Kedua, dilakukan tahap pra-pemrosesan teks yang mencakup proses pembersihan (cleansing), normalisasi, serta penghapusan duplikasi, guna memastikan bahwa data dalam kondisi siap untuk dikonversi menjadi representasi numerik. Ketiga, embedding berbasis LLM diterapkan pada judul-judul yang telah diproses, menggunakan model text-embedding-3-small dari OpenAI. Embedding ini mampu menangkap kedekatan makna antarjudul meskipun secara leksikal berbeda, dan dipilih karena kemampuannya merepresentasikan konteks semantik dalam teks pendek. Keempat, reduksi dimensi dan klasterisasi dilakukan dengan mengaplikasikan UMAP untuk menyederhanakan embedding berdimensi tinggi, yang kemudian dikelompokkan menggunakan algoritma HDBSCAN. Algoritma ini adaptif terhadap variasi bentuk dan ukuran klaster, serta mampu mengidentifikasi entri-entri yang tidak memiliki kedekatan semantik dengan kelompok manapun (noise). Terakhir, dilakukan interpretasi klaster dengan bantuan LLM generatif GPT-3.5-turbo, yang secara otomatis memberikan label tematik pada klaster melalui analisis isi seperti wordcloud atau kutipan judul. Hasil interpretasi ini digunakan untuk memetakan isu-isu dominan dalam pemberitaan politik mengenai Prabowo, serta memberikan pemahaman yang lebih mendalam terhadap struktur wacana publik selama periode pemilu.

Penelitian ini menggunakan pendekatan unsupervised learning untuk mengidentifikasi dan mengelompokkan isu-isu publik yang berkaitan dengan Presiden Prabowo berdasarkan judul-judul berita daring selama periode Pemilu 2024. Proses metode terdiri dari empat tahap utama: pengumpulan data, representasi teks (embedding), reduksi dimensi, dan klasterisasi.

Pengumpulan dan Praproses Data

Data berupa judul berita daring dikumpulkan dari berbagai portal berita nasional melalui metode scraping. Kriteria pengumpulan mencakup periode Januari hingga April 2024, dengan kata kunci seperti Prabowo, pemilu, Gibran, koalisi, dan kata kunci turunannya. Setelah dikumpulkan, dilakukan proses preprocessing seperti penghapusan duplikasi, penghapusan tanda baca dan angka yang tidak relevan, serta normalisasi teks ke format huruf kecil (lowercasing).

Embedding Teks dengan LLM

Setiap judul berita kemudian dikonversi menjadi vektor numerik berdimensi tinggi menggunakan model LLM (Large Language Model). Dalam hal ini, digunakan model text-embedding-3-small dari OpenAI yang dirancang khusus untuk menghasilkan representasi semantik dari teks pendek (Ye et al., 2024; Devlin et al., 2019). Model ini mengubah setiap judul menjadi vektor berdimensi 1536, yang menangkap konteks dan makna semantik dari teks secara lebih dalam dibanding metode klasik seperti TF-IDF.

Reduksi Dimensi Dengan UMAP

Karena dimensi embedding cukup tinggi, dilakukan proses reduksi dimensi menggunakan UMAP (Uniform Manifold Approximation and Projection) agar representasi dapat divisualisasikan dan dikelompokkan secara efisien. UMAP dipilih karena lebih unggul dibanding PCA dalam menjaga struktur lokal dan global data. Hasil reduksi umumnya direpresentasikan dalam 2–5 dimensi, tergantung pada kebutuhan visualisasi dan stabilitas klaster.

Klasterisasi dengan HDBSCAN

Proses klusterisasi dalam penelitian ini dilakukan dengan menggunakan algoritma HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), yang merupakan salah satu pendekatan klasterisasi berbasis densitas yang sangat sesuai untuk data teks pendek seperti judul berita. Keunggulan utama HDBSCAN terletak pada kemampuannya untuk bekerja tanpa memerlukan penentuan jumlah klaster (k) di awal, sehingga lebih fleksibel dibanding algoritma seperti KMeans. Selain itu, HDBSCAN mampu mengidentifikasi outlier, yakni data yang tidak cukup memiliki densitas untuk bergabung dalam klaster manapun—dalam konteks ini berarti judul-judul berita yang secara semantik terlalu unik atau tidak berhubungan dengan klaster dominan lainnya. Algoritma ini juga dapat menangani bentuk dan ukuran klaster yang beragam, sehingga lebih adaptif terhadap distribusi semantik dalam teks pendek.

Untuk menentukan nilai parameter utama seperti min_cluster_size dan min_samples, dilakukan pendekatan empiris melalui serangkaian eksperimen yang mengevaluasi sensitivitas hasil klasterisasi. Percobaan dilakukan dengan beberapa nilai min_cluster_size mulai dari 5 hingga 500. Hasilnya menunjukkan bahwa nilai terlalu kecil seperti 5 atau 10 menghasilkan jumlah klaster yang sangat besar (hingga 796 klaster), namun dengan ukuran rata-rata yang sangat kecil dan proporsi noise yang tetap tinggi. Sebaliknya, nilai yang terlalu besar seperti 200 atau 500 menyebabkan jumlah klaster menurun drastis dan sebagian besar data dikategorikan sebagai noise, yang berarti banyak isu tematik hilang dari representasi semantik.

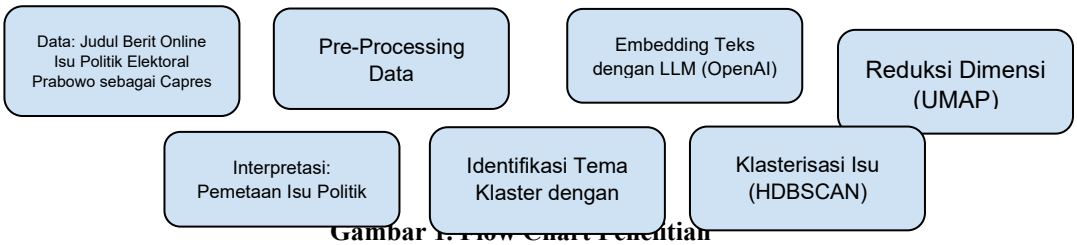
Nilai min_cluster_size = 50 dipilih sebagai konfigurasi optimal karena menghasilkan 85 klaster yang cukup tematik dengan proporsi noise yang realistis (27,2%), serta menjaga keseimbangan antara granularitas isu dan kohesi semantik dalam klaster. Pendekatan ini diharapkan dapat menghasilkan struktur klaster yang representatif terhadap keragaman isu publik yang beredar dalam pemberitaan tentang Presiden Prabowo selama periode Pemilu 2024.

Tabel 1. Perbandingan Data Jumlah Ukuran Klaster

min_cluster_size	n_clusters	n_noise	avg_cluster_size
5	796	12005	10.27
10	253	13651	25.81
20	103	14979	50.51
50	85	6527	20.56
100	3	11669	2.83
200	3	16445	1.24
500	2	15949	0.82

Pemberian Nama Klaster dengan LLM

Untuk membantu interpretasi klaster, penelitian ini juga menggunakan model gpt-3.5-turbo untuk memberikan nama atau label klaster secara otomatis. Beberapa judul dari setiap klaster dikirim sebagai prompt ke LLM, lalu model diminta memberikan ringkasan tema dalam 3–5 kata. Langkah ini bertujuan untuk memperkuat pemahaman naratif dari masing-masing klaster.

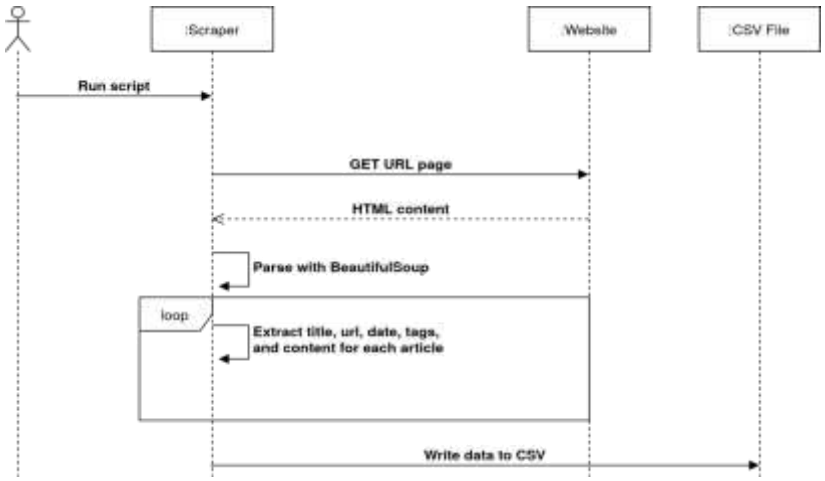


Gambar 1. Flow Chart Penelitian

HASIL DAN PEMBAHASAN

Pengambilan Data

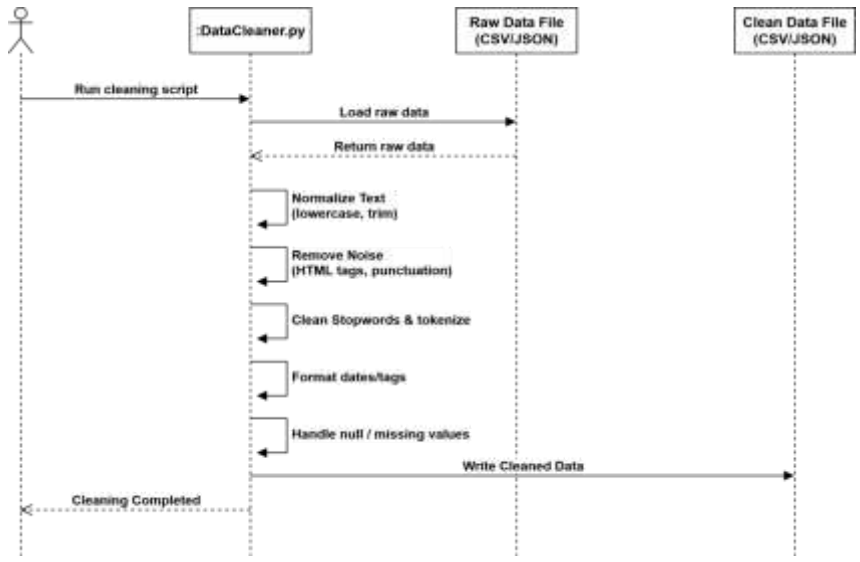
Pengambilan data dalam penelitian ini dilakukan secara sistematis dengan menggunakan teknik web scraping terhadap berbagai portal berita daring nasional. Fokus pengumpulan diarahkan pada judul-judul berita politik yang secara eksplisit maupun implisit menyebut nama Presiden Prabowo Subianto dalam konteks Pemilu 2024. Proses ini mencakup rentang waktu dari Januari hingga April 2024, untuk memastikan cakupan data yang cukup representatif terhadap dinamika wacana politik menjelang dan pasca pemilu. Kata kunci yang digunakan meliputi "Prabowo", "pemilu", "koalisi", "Gibran", serta variasi sinonim atau istilah terkait lainnya. Setelah pengumpulan, dilakukan penyaringan awal guna menghapus entri duplikat, mengeliminasi judul yang terlalu umum atau tidak relevan, serta menormalkan format teks ke dalam bentuk huruf kecil. Langkah ini bertujuan untuk menjamin kualitas dan konsistensi data sebelum masuk ke tahap representasi vektor menggunakan model embedding. Dengan total sekitar 24.000 judul berita yang berhasil dikumpulkan, korpus ini diharapkan mampu merepresentasikan spektrum isu politik yang muncul dalam diskursus publik digital seputar figur Prabowo Subianto selama periode elektoral.



Gambar 2. Sequence Diagram Pengambilan Data

Pra-pemrosesan Data

Tahap pra-pemrosesan data merupakan langkah krusial dalam memastikan kualitas korpus sebelum dilakukan analisis lanjutan berbasis embedding. Dalam penelitian ini, proses pembersihan data dilakukan terhadap seluruh judul berita yang telah dikumpulkan, dengan tujuan untuk menghilangkan elemen-elemen yang dapat mengganggu representasi semantik. Langkah awal melibatkan penghapusan duplikasi judul guna menghindari bias frekuensi yang dapat mempengaruhi proses klasterisasi. Selanjutnya, dilakukan pembersihan tanda baca, angka yang tidak bermakna, serta karakter non-alfabet yang tidak relevan dalam konteks wacana politik. Semua teks kemudian dinormalisasi ke format huruf kecil (lowercasing) untuk menyamakan bentuk kata dan meminimalisir redundansi leksikal. Proses ini juga disertai penghapusan stopwords dasar, jika diperlukan, untuk meningkatkan efisiensi embedding tanpa kehilangan makna kontekstual. Dengan demikian, hasil pra-pemrosesan ini menghasilkan korpus teks pendek yang bersih, konsisten, dan siap untuk dikonversi menjadi representasi numerik berdimensi tinggi melalui model LLM.



Gambar 3. Sequence Diagram Pra-pemrosesan Data

Embedding Token dengan LLMs

Setelah data melalui proses pra-pemrosesan, tahap selanjutnya adalah mengonversi judul berita menjadi representasi numerik menggunakan model Large Language Model (LLM). Dalam penelitian ini, digunakan model text-embedding-3-small dari OpenAI yang dirancang untuk menghasilkan embedding berdimensi tinggi (1536 dimensi) yang mampu menangkap kedekatan semantik antar teks pendek. Setiap judul berita dipetakan menjadi vektor yang tidak hanya mencerminkan makna literal, tetapi juga konteks yang tersembunyi dalam struktur kalimat. Proses ini memungkinkan analisis lanjutan seperti klasterisasi dilakukan berdasarkan kemiripan makna, bukan hanya kemiripan kata. Keunggulan model LLM dibandingkan pendekatan konvensional seperti TF-IDF atau word2vec terletak pada kemampuannya memahami hubungan konseptual dan sintaksis dalam teks yang sangat ringkas, seperti judul berita politik. Dengan demikian, embedding berbasis LLM menyediakan fondasi representasi yang lebih kaya dan kontekstual untuk mendeteksi pola narasi dalam diskursus media daring.

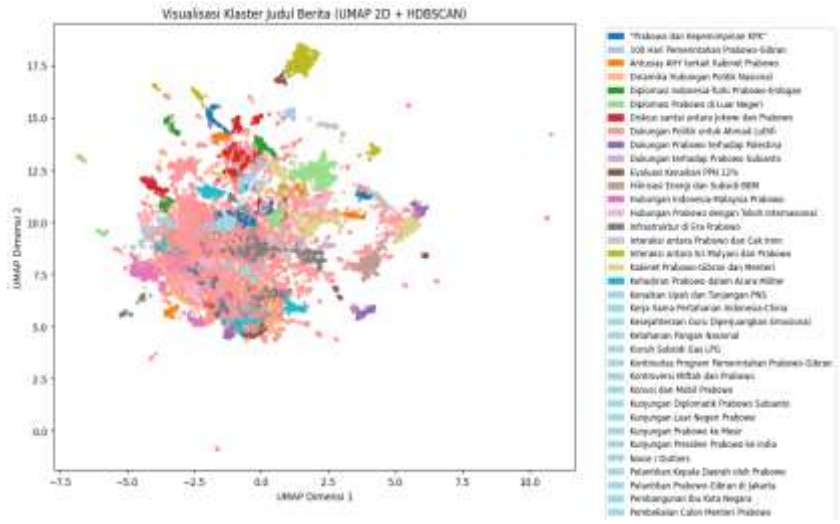
Gambar 4. Hasil Embedding Tokens dengan LLMs

Reduksi Dimensi (UMAP)

Setelah judul berita dikonversi menjadi vektor berdimensi tinggi melalui proses embedding menggunakan LLM, langkah berikutnya adalah mereduksi dimensi vektor tersebut agar lebih efisien untuk analisis visual dan klasterisasi. Dalam penelitian ini, digunakan algoritma Uniform Manifold Approximation and Projection (UMAP), sebuah teknik reduksi dimensi non-linear yang mampu mempertahankan struktur lokal dan global dari data asli. UMAP dipilih karena terbukti lebih unggul dibanding metode reduksi tradisional seperti PCA, terutama dalam menangani data teks yang kompleks dan berdimensi tinggi. Dengan mereduksi dimensi dari 1536 menjadi dua atau tiga dimensi, representasi data menjadi lebih mudah divisualisasikan dan lebih ringan secara komputasional, tanpa mengorbankan informasi semantik yang penting. Hasil reduksi ini digunakan sebagai masukan dalam proses klasterisasi dengan HDBSCAN, serta untuk membuat visualisasi peta sebaran isu politik yang terkandung dalam judul berita daring. Dengan demikian, UMAP berperan penting sebagai penghubung antara representasi semantik berbasis LLM dan struktur tematik yang dapat ditafsirkan secara visual dan analitis.

Hasil Proses Klasterisasi

Pada bagian ini dapat diuraikan mengenai hasil dari penelitian serta pengujian yang telah dilakukan. Selain itu, disampaikan juga mengenai pembahasan dari penelitian maupun pengujian yang telah dilakukan. Hasil dan pembahasan seharusnya merupakan bab yang paling banyak isinya pada sebuah paper. Isi Hasil dan Pembahasan dapat mencapai 50-65% dari keseluruhan paper.

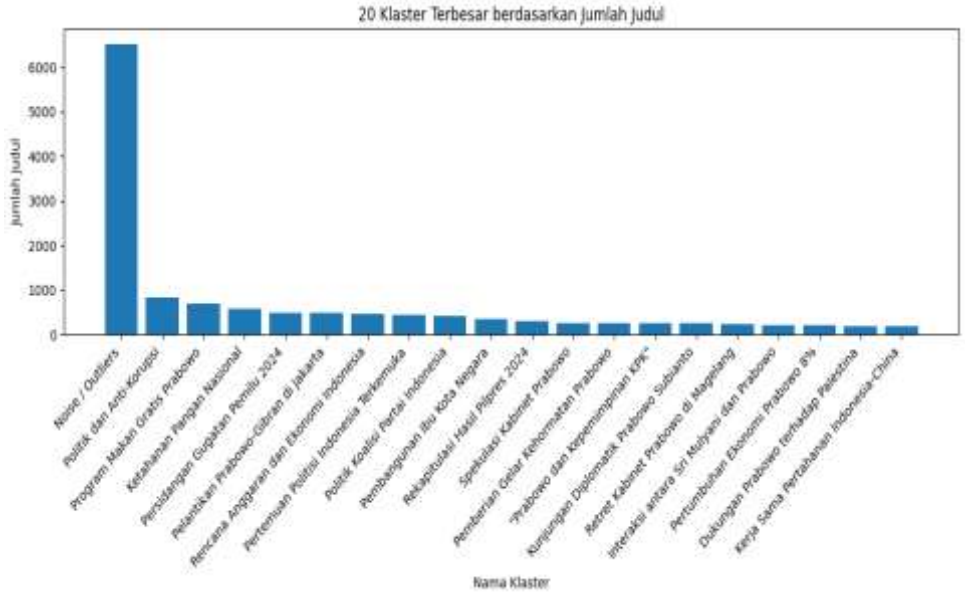


Gambar 5. Visualisasi Klaster Judul Berita (UMAD 2D + HDBSCAN)

Plot UMAP-HDBSCAN menampilkan persebaran 24 000 judul berita dalam ruang dua dimensi yang diproyeksikan dari embedding LLM. Titik-titik berwarna menggambarkan 85 klaster isu, sedangkan titik abu-abu (jika ada) merepresentasikan noise. Terlihat sebuah inti padat di pusat plot: di sini belasan klaster—misalnya "Kontroversi politik nasional", "Kenaikan upah dan tunjangan PNS", atau "Pertemuan Jokowi–Prabowo"—bertumpuk rapat. Hal ini menandakan bahwa judul-judul di dalamnya saling beririsan secara semantik; tema-tema tersebut menjadi arus utama pemberitaan sehingga embedding-nya cenderung berdekatan. Di sekitar inti muncul pulau-pulau klaster terpisah. Klaster seperti "Diplomasi Indonesia–Turki Prabowo–Erdogan" (bagian kiri-atas) atau "Krisis subsidi gas LPG" (kanan-atas) membentuk gugus tersendiri yang relatif jauh dari pusat, menandakan isu yang lebih spesifik dan leksikalnya khas, sehingga jaringan maknanya tidak bercampur dengan klaster umum. Beberapa klaster kecil lain—misalnya "Ucapan Selamat Kemenangan Prabowo–Gibran"—muncul sebagai gugus sempit di tepi plot, menunjukkan konsistensi konteks yang kuat meskipun frekuensi judulnya tidak besar.

Sebaran warna yang cukup merata memperlihatkan bahwa HDBSCAN berhasil mendeteksi keragaman topik tanpa memaksa jumlah klaster tertentu. Namun, tumpang-tindih kecil antar warna di pusat plot juga mengindikasikan overlap semantik; beberapa isu—contohnya "Penataan Kabinet" dan "Spekulasi Kabinet Prabowo"—berdekatan karena kosa kata dan framing-nya mirip. Secara keseluruhan, visual ini mengonfirmasi bahwa kombinasi embedding LLM + UMAP + HDBSCAN efektif memetakan lanskap wacana: isu-isu dominan terkonsentrasi di pusat, sedangkan tema yang sangat spesifik atau kurang populer muncul sebagai gugus pinggiran atau menjadi noise.

Interpretasi Tema dan Narasi Klaster



Gambar 6. Dua puluh klaster terbesar berdasarkan jumlah judul

Plot UMAP-HDBSCAN menampilkan persebaran 24 000 judul berita dalam ruang dua dimensi yang diproyeksikan dari embedding LLM. Titik-titik berwarna menggambarkan 85 klaster isu, sedangkan titik abu-abu (jika ada) merepresentasikan noise. Terlihat sebuah inti padat di pusat plot: di sini belasan klaster—misalnya "Kontroversi politik nasional", "Kenaikan upah dan tunjangan PNS", atau "Pertemuan Jokowi–Prabowo"—bertumpuk rapat. Hal ini menandakan bahwa judul-judul di dalamnya saling beririsan secara semantik; tema-tema tersebut menjadi arus utama pemberitaan sehingga embedding-nya cenderung berdekatan.Di sekitar inti muncul pulau-pulau klaster terpisah. Klaster seperti "Diplomasi Indonesia–Turki Prabowo–Erdogan" (bagian kiri-atas) atau "Krisis subsidi gas LPG" (kanan-atas) membentuk gugus tersendiri yang relatif jauh dari pusat, menandakan isu yang lebih spesifik dan leksikalnya khas, sehingga jaringan maknanya tidak bercampur dengan klaster umum. Beberapa klaster kecil lain—misalnya "Ucapan Selamat Kemenangan Prabowo–Gibran"—muncul sebagai gugus sempit di tepi plot, menunjukkan konsistensi konteks yang kuat meskipun frekuensi judulnya tidak besar.

Sebaran warna yang cukup merata memperlihatkan bahwa HDBSCAN berhasil mendeteksi keragaman topik tanpa memaksa jumlah klaster tertentu. Namun, tumpang-tindih kecil antarwarna di pusat plot juga mengindikasikan overlap semantik; beberapa isu—contohnya "Penataan Kabinet" dan "Spekulasi Kabinet Prabowo"—berdekatan karena kosa kata dan framing-nya mirip. Secara keseluruhan, visual ini mengonfirmasi bahwa kombinasi embedding LLM + UMAP + HDBSCAN efektif memetakan lanskap wacana: isu-isu dominan terkonsentrasi di pusat, sedangkan tema yang sangat spesifik atau kurang populer muncul sebagai gugus pinggiran atau menjadi noise.

Tabel 2. Tema Berita

Judul Berita
Pratikno: Tidak Ada Seskab-Menaker Definitif hingga Prabowo Dilantik Presiden
Jokowi dan Gibran Dampingi Prabowo di HUT ke-17 Gerindra
Prabowo Absen di Pelantikan Bahlil, Supratman, Rosan hingga Hasan
Tujuh Duta Besar Negara Berikan Surat Kepercayaan ke Presiden Prabowo
Minta Presiden Dipilih MPR Lagi, La Nyalla Desak Sidang Istimewa Usai Prabowo Dilantik

Catat Kepuasan Tertinggi ke Prabowo, PAN: Kami Loyal dan Setia Mendukung
5 Jaksa Agung Muda hingga Kepala PPATK Merapat ke Istana Bertemu Prabowo
Zulkifli Hasan Ingatkan 9 Kader PAN di Kabinet Prabowo Wujudkan Swasembada Pangan
Saat Revisi UU Kementerian Negara Akan Jadi Acuan Prabowo Susun Kabinet, Pembahasannya Disebut Kebetulan.

Pada hasil klasterisasi, sejumlah judul berita tidak tergabung ke dalam klaster-klaster tematik utama dan dikategorikan sebagai noise. Berdasarkan sampel terhadap 20 judul berita dari klaster noise, ditemukan bahwa mayoritas judul tersebut memiliki karakteristik yang cenderung umum, fragmentaris, atau sangat pendek sehingga tidak memiliki kekuatan semantik yang cukup untuk dikelompokkan. Banyak judul hanya memuat informasi dasar atau penyebutan tokoh politik tanpa narasi yang jelas, sehingga embedding yang dihasilkan tidak menunjukkan kedekatan makna yang signifikan dengan klaster-klaster yang terbentuk. Beberapa judul juga memperlihatkan variasi topik yang tinggi atau menggunakan gaya bahasa yang ambigu, seperti clickbait, yang semakin memperumit proses pengelompokan berbasis semantik.

Keberadaan klaster noise dalam hasil ini justru memperkuat efektivitas metode yang digunakan, karena algoritma HDBSCAN mampu secara adaptif mengidentifikasi data yang tidak memenuhi kriteria kepadatan tematik. Dalam konteks analisis teks pendek, fenomena noise adalah hal yang wajar mengingat heterogenitas wacana publik di media daring. Selain itu, isolasi terhadap judul-judul yang bersifat noise membantu menjaga kualitas dan koherensi klaster utama, sehingga interpretasi terhadap struktur isu publik yang berkaitan dengan Presiden Prabowo dapat dilakukan dengan lebih akurat dan bermakna. Dengan demikian, keberadaan noise bukan hanya mencerminkan keterbatasan data, melainkan juga memperlihatkan sensitivitas model dalam menangkap kompleksitas semantik teks politik elektoral di Indonesia.

Pada hasil klasterisasi, sebanyak 6.527 judul berita (27,2% dari total korpus) tidak tergabung ke dalam klaster-klaster tematik utama dan dikategorikan sebagai noise. Berdasarkan sampel terhadap 20 judul dari kategori ini, mayoritas memiliki karakteristik yang cenderung umum, fragmentaris, atau sangat pendek sehingga tidak memiliki kekuatan semantik yang cukup untuk dikelompokkan. Analisis statistik menunjukkan bahwa rata-rata panjang judul noise adalah 10,54 kata (63,52 karakter), hampir setara dengan rata-rata panjang judul dalam klaster tematik (10,70 kata dan 64,37 karakter). Hal ini mengindikasikan bahwa panjang teks bukanlah faktor utama yang menyebabkan suatu judul menjadi noise. Faktor yang lebih dominan adalah rendahnya kekayaan leksikal dan kurangnya konsistensi topik. Kata-kata yang paling sering muncul dalam kategori ini adalah istilah generik seperti “Prabowo”, “Gibran”, “Presiden”, dan “Menteri”, yang tidak cukup spesifik untuk membentuk kedekatan semantik yang kuat di ruang embedding.

Banyak judul noise hanya memuat informasi dasar atau penyebutan tokoh politik tanpa narasi tematik yang jelas, misalnya “Pratikno: Tidak Ada Seskab-Menaker Definitif hingga Prabowo Dilantik Presiden” atau “Jokowi dan Gibran Dampingi Prabowo di HUT ke-17 Gerindra”. Beberapa judul lainnya menampilkan variasi topik yang tinggi atau gaya bahasa ambigu seperti clickbait, sehingga semakin menyulitkan proses pengelompokan berbasis semantik. Kondisi ini mencerminkan dinamika pemberitaan politik digital pada periode Pemilu 2024 yang sarat dengan wacana reaktif, sensasional, dan kurang terstruktur secara tematik. Keberadaan proporsi noise yang relatif tinggi justru menunjukkan sensitivitas algoritma HDBSCAN dalam mengisolasi data yang tidak memenuhi kriteria kepadatan tematik, sehingga kualitas dan koherensi klaster utama tetap terjaga. Dalam konteks analisis teks pendek seperti judul berita politik, fenomena noise merupakan hal yang wajar mengingat tingginya

heterogenitas wacana publik di media daring. Dengan memisahkan judul-judul yang bersifat noise, interpretasi terhadap struktur isu publik yang berkaitan dengan Presiden Prabowo dapat dilakukan dengan lebih akurat, sekaligus memperlihatkan kemampuan model dalam menangkap kompleksitas semantik teks politik elektoral di Indonesia.

Pembahansan

Setelah penerapan metode HDBSCAN untuk mengelompokkan judul berita daring mengenai isu politik elektoral di Indonesia, embedding yang dihasilkan oleh model bahasa besar (LLMs) berfungsi signifikan dalam menggambarkan serta menganalisis klaster-klaster tematik tersebut. HDBSCAN, yang dikenal karena kemampuannya dalam menangani berbagai kepadatan dalam data, sangat sesuai untuk mengklasifikasikan berita yang memiliki tema beragam dan kompleks, seperti dalam konteks politik elektoral (Neto et al., 2017). Proses analisis ini dimulai dengan pembentukan embedding dari judul berita menggunakan LLM, yang mampu menangkap nuansa semantik dan latar konteks teks (Suchanek & Tuan, 2023). Embedding yang dihasilkan membawa kesadaran kontekstual yang tinggi terhadap isi berita, memungkinkan pembentukan representasi yang kaya secara makna. Setelah proses klasterisasi dengan HDBSCAN, setiap klaster yang terbentuk dapat dianggap sebagai kelompok berita yang memiliki kesamaan tematik. Misalnya, satu klaster dapat merepresentasikan isu "kecurangan dalam pemilu", sementara klaster lain berfokus pada "dinasti politik". Penalaran semantik yang mendalam ini memungkinkan identifikasi bukan hanya kata kunci, tetapi juga hubungan logis, pola makna, serta konteks sosial yang mendasarinya (He et al., 2024).

Dalam konteks representasi politik digital, penting untuk mempertimbangkan juga bagaimana narasi terbentuk di luar institusi formal (Fossheim, 2022). Dalam lanskap komunikasi kontemporer, media digital memungkinkan aktor non-elektoral---seperti jurnalis, warganet, atau figur publik---untuk turut serta membentuk makna politik melalui penyebaran narasi, simbol, dan framing isu. Representasi politik tidak lagi hanya dimediasi oleh lembaga resmi seperti partai atau parlemen, tetapi juga oleh struktur wacana yang terbentuk secara organik di ruang daring. Hal ini menyebabkan legitimasi naratif menjadi semakin kompleks dan terfragmentasi, di mana isu-isu seperti pemilu, kebijakan, atau kontroversi tokoh dimaknai dan dinegosiasikan ulang melalui praktik diskursif yang tersebar luas. Oleh karena itu, pemetaan klaster isu tidak hanya merekam dinamika semantik, tetapi juga mencerminkan konfigurasi aktor sosial dan politik yang membentuk medan representasi secara kolektif dan desentralistik.

Analisis lebih lanjut terhadap masing-masing klaster memberikan pemahaman mengenai frekuensi, kepadatan, dan hubungan antar judul berita dalam klaster. Informasi ini dapat dimanfaatkan untuk mengungkap pola penyebaran isu dan sensitivitas tema tertentu menjelang periode kritis seperti pemilu. Misalnya, topik tentang "kecurangan pemilu" dapat menunjukkan lonjakan jumlah berita saat mendekati hari pemilihan, sebagaimana terlihat dari analisis ukuran klaster.

Dengan demikian, melalui integrasi embedding berbasis LLM dan klasterisasi menggunakan HDBSCAN, representasi judul berita tidak lagi dipandang hanya sebagai sekumpulan teks terpisah, melainkan sebagai entitas semantik yang mencerminkan dinamika sosial dan politik. Pendekatan ini menawarkan alat analisis yang berharga untuk mengeksplorasi interaksi antara media, wacana publik, dan perubahan politik di Indonesia dalam konteks Pemilu 2024. Hasil penelitian ini menunjukkan bahwa kombinasi antara embedding berbasis LLM dan algoritma klasterisasi HDBSCAN mampu mengelompokkan judul berita politik elektoral secara efektif berdasarkan kedekatan semantik. Temuan ini sejalan dengan penelitian yang dilakukan oleh Keraghel et al. (2024) dan studi NLP transformasional seperti BERT dan T5 yang menunjukkan kemampuan tinggi dalam pemahaman konteks

(Devlin et al., 2019; Raffel et al., 2020). Kondisi ini menunjukkan bahwa embedding berbasis model bahasa besar dapat menangkap hubungan semantik yang halus dalam teks pendek, sehingga meningkatkan akurasi klasterisasi dibandingkan dengan representasi berbasis frekuensi kata tradisional seperti TF-IDF. Penerapan embedding dari OpenAI dalam penelitian ini memperkuat bukti empiris bahwa teknik representasi semantik modern lebih unggul dalam mengelola data berita yang heterogen dan kontekstual.

Dari sisi teknik klasterisasi, penggunaan HDBSCAN terbukti efektif dalam mengidentifikasi struktur isu tanpa perlu menentukan jumlah klaster di awal, sesuai dengan rekomendasi Weng et al. (2022). HDBSCAN berhasil mendeteksi 85 klaster tematik serta mengisolasi sekitar 27,2% data sebagai noise, sebuah hasil yang mencerminkan fleksibilitas algoritma ini dalam menangani distribusi data yang tidak merata, sebagaimana juga dijelaskan dalam studi sebelumnya. Tingginya proporsi noise yang teridentifikasi dalam penelitian ini tidak menunjukkan kegagalan model, melainkan mencerminkan kenyataan fragmentasi wacana di media daring politik, yang juga ditemukan dalam penelitian terkait oleh Yang et al. (2024). Selain itu, penggunaan UMAP sebagai teknik reduksi dimensi sebelum klasterisasi mendukung efektivitas pipeline analisis ini, sebagaimana diusulkan oleh Blanco-Portals et al. (2021). UMAP memungkinkan pelestarian struktur lokal dan global data embedding, sehingga proses klasterisasi HDBSCAN dapat berjalan lebih optimal di ruang berdimensi lebih rendah tanpa kehilangan kedekatan semantik penting antar judul berita. Secara umum, hasil penelitian ini mengonfirmasi bahwa integrasi embedding berbasis LLM, teknik reduksi dimensi UMAP, dan klasterisasi adaptif HDBSCAN membentuk pendekatan yang robust dan scalable untuk analisis wacana digital, terutama dalam konteks teks pendek dan dinamis seperti judul berita politik elektoral. Temuan ini memperkuat literatur eksisting dan menawarkan landasan empiris bagi penelitian lanjutan di bidang analisis media digital dan politik berbasis data besar.

Kesimpulan

Penelitian ini bertujuan untuk mengelompokkan judul-judul berita politik daring yang berkaitan dengan Presiden Prabowo Subianto selama periode Pemilu 2024, dengan menggunakan pendekatan berbasis embedding Large Language Models (LLMs) dari OpenAI dan algoritma klasterisasi HDBSCAN. Berdasarkan analisis terhadap 24.000 judul berita, ditemukan bahwa kombinasi embedding semantik dan klasterisasi adaptif mampu mengidentifikasi sebanyak 85 klaster tematik serta mengisolasi sekitar 27,2% data sebagai noise. Embedding berbasis LLM berhasil menangkap kedekatan semantik antar teks pendek dengan lebih akurat dibandingkan metode representasi konvensional, sebagaimana didukung oleh penelitian terdahulu. Sementara itu, HDBSCAN terbukti efektif dalam menangani keragaman densitas data, tanpa memerlukan penentuan jumlah klaster secara eksplisit, serta mampu mengidentifikasi data outlier yang tidak konsisten secara semantik. Visualisasi hasil klasterisasi melalui UMAP menunjukkan bahwa sebagian besar klaster terdistribusi dengan baik, dengan beberapa kelompok isu utama seperti elektabilitas, politik dinasti, ketahanan pangan, dan dinamika pasca pemilu yang menonjol dalam struktur wacana publik. Selain menghasilkan pemetaan isu yang komprehensif, penelitian ini juga mengungkap fragmentasi wacana politik digital di Indonesia, sebagaimana tercermin dalam tingginya proporsi noise. Hal ini memperlihatkan dinamika pemberitaan yang heterogen dan memperkuat relevansi pendekatan unsupervised berbasis LLM dan HDBSCAN dalam studi analisis media. Penelitian ini berkontribusi dalam memperkaya metode analisis wacana digital berbasis data besar, khususnya dalam konteks politik elektoral. Temuan-temuan yang diperoleh tidak hanya memperkuat literatur yang ada, tetapi juga membuka ruang untuk penelitian lanjutan, seperti pelacakan evolusi isu dari waktu ke waktu, analisis sentimen berbasis klaster, serta pemetaan aktor digital dalam membentuk narasi publik.

DAFTAR PUSTAKA

Blanco-Portals, J., Peiró, F., & Estradé, S. (2021). Strategies for EELS data analysis: Introducing UMAP and HDBSCAN for dimensionality reduction and clustering. *Microscopy and Microanalysis, 28*(1), 109–122. https://doi.org/10.1017/s1431927621013696

Campello, R., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data, 10*(1), 1–51. https://doi.org/10.1145/2733381

Chajia, M., & Nfaoui, E. (2024). Customer churn prediction approach based on LLM embeddings and logistic regression. *Future Internet, 16*(12), 453. https://doi.org/10.3390/fi16120453

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*. https://doi.org/10.48550/arXiv.1810.04805

Fossheim, K. (2022). How can non-elected representatives secure democratic representation? *Policy & Politics, 50*(2), 243–260. https://doi.org/10.1332/030557321x16371011677734

He, H., Du, C., Fu, K., Wen, B., Sun, Y., Peng, J., & Chang, L. (2024). Human-like object concept representations emerge naturally in multimodal large language models. *Research Square*. https://doi.org/10.21203/rs.3.rs-4641719/v1

Hidayat, M. (2024). The 2024 general elections in Indonesia: Issues of political dynasties, electoral fraud, and the emergence of a national protest movement. *IASJOL, 2*(1), 33–51. https://doi.org/10.62033/iasjol.v2i1.51

Keraghel, I., Morbieu, S., & Nadif, M. (2024). Beyond words: A comparative analysis of LLM embeddings for effective clustering. In *Proceedings* (pp. 205–216). https://doi.org/10.1007/978-3-031-58547-0_17

Korade, N., Salunke, M., Bhosle, A., Asalkar, G., Lal, B., & Kumbharkar, P. (2025). Elevating intelligent voice assistant chatbots with natural language processing and OpenAI technologies. *Indonesian Journal of Electrical Engineering and Computer Science, 37*(1), 507–517. https://doi.org/10.11591/ijeecs.v37.i1.pp507-517

Lewis, P., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*. https://doi.org/10.48550/arXiv.1910.13461

Motoki, F., Neto, V., & Rodrigues, V. (2023). More human than human: Measuring ChatGPT political bias. *Public Choice, 198*(1–2), 3–23. https://doi.org/10.1007/s11127-023-01097-2

Neto, A., Sander, J., Campello, R., & Nascimento, M. (2017). Efficient computation of multiple density-based clustering hierarchies. In *Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM)*. https://doi.org/10.1109/icdm.2017.127

Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*, 1–67. https://doi.org/10.48550/arXiv.1910.10683

Sadeghi, S., Bui, A., Forooghi, A., Lü, J., & Ngom, A. (2024). Can large language models understand molecules? *BMC Bioinformatics, 25*(1). https://doi.org/10.1186/s12859-024-05847-x

Suchanek, F., & Tuan, L. (2023). Knowledge bases and language models: Complementing forces. In *Proceedings* (pp. 3–15). https://doi.org/10.1007/978-3-031-45072-3_1

Weng, M., Wu, S., & Dyer, M. (2022). Identification and visualization of key topics in scientific publications with transformer-based language models and document clustering methods. *Applied Sciences, 12*(21), 11220. https://doi.org/10.3390/app122111220

Yang, C., Cao, B., & Fan, J. (2024). TEC: A novel method for text clustering with large language models guidance and weakly-supervised contrastive learning. *Proceedings of the International AAAI Conference on Web and Social Media, 18*, 1702–1712. https://doi.org/10.1609/icwsm.v18i1.31419

Ye, Z., Ai, Q., Liu, Y., de Rijke, M., Zhang, M., Lioma, C., & Ruotsalo, T. (2024). Generative language reconstruction from brain recordings. *Research Square*. https://doi.org/10.21203/rs.3.rs-4587150/v1