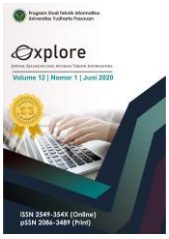
Halaman jurnal: http://jurnal.yudharta.ac.id/v2/index.php/EXPLORE-IT/

Explore IT! Jurnal Keilmuan dan Aplikasi Teknik Informatika

Akreditasi Sinta 5 (S5) – SK No. 36/E/KPT/2019
p-ISSN 2086-3489 (Print)- e-ISSN 2549-354X (Online)
DOI https://doi.org/10.35891/explorit



Klasterisasi Berita Berbahasa Indonesia Menggunakan Model K-Means Dan Indobert Untuk Menentukan Berita Hoaks

Clustering Indonesian-Language News Using the K-Means Model and Indobert to Determine Hoax News

Kholis Anwari [1], Totok Chamidy [2], Suhartono [3]

[1,2,3] Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang

email: [1] 210605210012@student.uin-malang.ac.id, [2] to2k2013@ti.uin-malang.ac.id, [3] suhartono@ti.uin-malang.ac.id,

INFO ARTIKEL

Sejarah Artikel:
Diterima 27 Juni 2025
Direvisi 28 Juni 2025
Disetujui 29 Juni 2025
Dipublikasi 30 Juni 2025

Katakunci:
IndoBERT
K-Means Clustering
Berita Hoaks

ABSTRAK

Penyebaran berita hoaks di media digital, khususnya dalam konteks politik, semakin marak dan berdampak negatif terhadap persepsi publik serta stabilitas sosial. Penelitian ini bertujuan untuk menganalisis dan mengelompokkan berita politik berbahasa Indonesia menggunakan model IndoBERT sebagai representasi teks dan algoritma K-Means Clustering sebagai metode klasterisasi, guna membedakan berita hoaks dan fakta. IndoBERT digunakan untuk menghasilkan vektor representasi kata berbasis konteks, vektor ini kemudian dikelompokkan menggunakan K-Means Clustering untuk membentuk klaster berdasarkan kemiripan semantik. Hasil penelitian menunjukkan bahwa pendekatan ini berhasil memisahkan berita ke dalam dua klaster utama, yaitu hoaks dan fakta, dengan hasil evaluasi kuantitatif berupa nilai Silhouette Score sebesar 0,21. Metrik ini mengukur seberapa baik suatu data cocok dengan klasternya sendiri dibandingkan dengan klaster lainnya. Nilai Silhouette Score berada pada rentang –1 hingga +1. Validasi manual terhadap 200 sampel berita menunjukkan bahwa klaster yang terbentuk memiliki tingkat akurasi klasifikasi sebesar 88% untuk berita fakta dan 82% untuk berita hoaks. Temuan ini membuktikan bahwa kombinasi IndoBERT dan K-Means Clustering efektif digunakan sebagai alat bantu dalam identifikasi berita hoaks secara otomatis dan unsupervised. Penelitian ini juga merekomendasikan penggunaan dataset yang lebih luas dan penggabungan dengan pendekatan supervised untuk hasil yang lebih optimal.

Keyword:
IndoBERT
K-Means Clustering Fake News

DOI Artikel:
https://doi.org/10.35891/explorit.v17i1.6502

ABSTRACT

Abstract *– The spread of fake news in digital media, especially in the political context, is becoming increasingly prevalent and has a negative impact on public perception and social stability. This study aims to analyze and classify Indonesian-language political news using the IndoBERT model as text representation and the K-Means Clustering algorithm as a clustering method to distinguish between fake news and factual news. IndoBERT is used to generate context-based word representation vectors, which are then clustered using K-Means Clustering to form clusters based on semantic similarity. The results show that this approach successfully separates news into two main clusters, namely hoaxes and facts, with a quantitative evaluation result of a Silhouette Score of 0.21. This metric measures how well a data point fits into its own cluster compared to other clusters. The Silhouette Score ranges from -1 to +1. Manual validation of 200 news samples showed that the formed clusters had a classification accuracy rate of 88% for factual news and 82% for hoax news. These findings prove that the combination of IndoBERT and K-Means Clustering is effective as a tool for automatically and unsupervised identification of hoax news. This study also recommends the use of a broader dataset and combination with a supervised approach for more optimal results.*

1. Pendahuluan

Peran internet dan teknologi di masyarakat semakin pesat, terutama dengan semakin banyaknya berita yang beredar dan mudah didapatkan dalam hitungan detik. Selain itu, teknologi telah membuat penyebaran berita menjadi lebih masif dengan menyediakan berbagai saluran seperti media sosial, website, blog, saluran serta situs-situs di mana orang dapat membaca berita terbaru kapanpun dan dimanapun mereka mau. Sangat mungkin mereka untuk tetap mendapatkan berita terbaru dan memberikan kesempatan untuk mendiskusikan topik-topik yang sedang hangat dengan cepat [1].

Internet merupakan irisan yang tidak terpisahkan dari kehidupan sebagian besar penduduk dunia. Perkembangan dan adopsi internet yang sangat cepat telah mendorong laju pertumbuhan serta pertukaran informasi secara signifikan dibandingkan dengan era sebelumnya, sehingga jumlah informasi mengalami peningkatan drastis [2].

Penyebaran berita hoaks merupakan tantangan serius di era digital, terutama di Indonesia yang memiliki jumlah pengguna internet dan media sosial yang sangat besar. Informasi yang salah atau menyesatkan dapat menyebar secara masif dan sulit dikendalikan, sehingga menimbulkan dampak signifikan terhadap berbagai aspek kehidupan masyarakat, mulai dari sosial, budaya, hingga politik.Dari segi sosial, hoaks dapat memicu kepanikan, kebingungan, bahkan konflik horizontal antar kelompok masyarakat. Hal ini dapat melemahkan partisipasi demokratis dan menciptakan persepsi publik yang bias terhadap realitas sosial-politik [3].

Meningkatnya jumlah berita digital menimbulkan tantangan bagi pembaca dalam menemukan informasi yang relevan secara cepat dan efisien. Proses pencarian berita masih dilakukan secara umum, yang tidak hanya memakan waktu tetapi juga membutuhkan sumber daya yang besar. Selain itu, mesin pencari seperti search engine pun kesulitan dalam menyajikan berita-berita yang saling berkaitan atau memiliki kesamaan topik secara otomatis. Masalah ini semakin kompleks karena data berita umumnya tidak disertai label atau kategori yang jelas. Akibatnya, klasifikasi dan pengorganisasian berita menjadi kurang optimal. Dimensi data yang tinggi dalam teks berita juga menjadi hambatan tersendiri dalam proses pengolahan informasi secara efektif. Dengan demikian, diperlukan suatu solusi yang mampu mengelompokkan berita ke dalam kelompok-kelompok yang serupa berdasarkan isi, fakta, atau topik pembahasannya. Pengelompokan ini dapat dilakukan tanpa memerlukan label sebelumnya, sehingga metode seperti klasterisasi menjadi pendekatan yang tepat untuk mengatasi permasalahan tersebut [4].

Klasterisasi berita merupakan salah satu pendekatan dalam bidang data mining yang bertujuan untuk pengelompokkan berita dari berbagai sumber ke dalam klaster yang memiliki kesamaan tertentu. Setiap klaster berisi kumpulan berita yang membahas topik yang serupa atau berkaitan erat, sedangkan berita yang berada di klaster berbeda umumnya memiliki tingkat kesamaan topik yang rendah atau tidak berkaitan [5]. Tingkat kemiripan (similarity) antar berita dalam satu klaster serta tingkat ketidakmiripan (dissimilarity)antar klaster menjadi indikator utama dalam menilai kualitas hasil klasterisasi yang dilakukan. Semakin tinggi kesamaan dalam satu klaster dan semakin besar perbedaan antar klaster, maka semakin baik hasil klasterisasi tersebut. Namun, jumlah data berita yang sangat besar dan terus bertambah dapat memengaruhi performa dari proses klasterisasi itu sendiri. Oleh sebab itu, diperlukan proses yang sistematis mulai dari tahap ekstraksi fitur, representasi data, hingga pembangunan model klasterisasi yang efektif agar diperoleh hasil pengelompokan yang akurat dan efisien.

Salah satu pendekatan untuk membedakan antara berita hoaks dan berita faktual, serta untuk mencegah penyebaran informasi secara sembarangan, adalah melalui proses klasterisasi berita. Proses ini bertujuan untuk membantu masyarakat dalam membedakan dengan jelas antara berita faktual dan berita hoaks. Dengan demikian, masyarakat dapat lebih waspada dan tidak mudah ikut arus untuk mempercayai, serta menyebarkan informasi yang belum terverifikasi kebenarannya, agar tidak memunculkan kesalahpahaman di tengah masyarakat serta mencegah tersebarnya informasi yang bersifat menyesatkan [6]. Tindakan ini merupakan salah satu bentuk tanggung jawab moral sekaligus langkah yang baik dalam mencegah penyebaran informasi hoaks. Penyebaran informasi tanpa melalui proses klarifikasi tidak hanya berpotensi menimbulkan kesalahpahaman, tetapi juga dapat menciptakan keresahan di tengah masyarakat. Oleh karena itu, sikap selektif dan teliti dalam menyaring informasi menjadi bagian penting dari upaya membangun ekosistem komunikasi yang sehat, jujur, dan dapat dipercaya, baik dalam konteks sosial maupun dalam ruang digital yang terbuka.

Salah satu model yang efektif adalah IndoBERT (Bidirectional Encoder Representation from Transformer), pemilihan model IndoBERT berfokus pada pengoptimalan hyperparameter yang didasarkan pada relevansi dan kekuatan model dalam memahami konteks teks berbahasa Indonesia. IndoBERT mampu menangkap makna kata-kata dalam bahasa lokal secara mendalam, menjadikannya sangat tepat dan terukur untuk mendeteksi teks informal di media online.

Penelitian deteksi berita hoaks telah banyak dilakukan, salah satunya dengan pendekatan klasterisasi berbasis IndoBERT. Tantangan umum yang dihadapi dalam proses klasterisasi adalah data yang sangat banyak. Pemanfaatkan IndoBERT untuk mendeteksi hoaks politik, data terdiri dari 20.928 berita faktual [7]. Penelitian Implementasi metode K-Means Clustering untuk peringkasan teks ekstraktif, di mana kalimat direpresentasikan sebagai vektor dalam ruang dimensi berdasarkan vocabulary dokumen. Metode ini berhasil mengelompokkan kalimat berdasarkan tingkat kesamaan semantik, memungkinkan pemilihan kalimat representatif dari setiap cluster untuk membentuk ringkasan akhir. Pendekatan ini memastikan bahwa ringkasan yang dihasilkan mencakup informasi penting dari berbagai topik dalam dokumen [8].

Menurut penelitian lain K-Means Clustering merupakan teknik pengelompokan data yang memisahkan dataset menjadi beberapa kelompok berbeda, di mana setiap kelompok memiliki titik pusat atau centroid tersendiri sebagai representasi karakteristik kelompok tersebut [9]Tujuan dari K-Means yaitu untuk memaksimalkan variasi antar kelompok dan suatu kelompok secara keseluruhan.

Efektivitas metode ini telah dibuktikan melalui berbagai penelitian untuk pelanggan potensial [10], untuk menganalisis pengelompokan data bongkar muat di Provinsi Riau [11]. Temuan tersebut menunjukkan bahwa metode ini bersifat fleksibel dan dapat dengan mudah diadaptasi untuk berbagai jenis data.

Berdasarkan permasalahan yang telah dijelaskan diatas, maka disusunlah penelitian terkait bagaimana pengelompokan berita dengan menggunakan konsep clustering untuk mengetahui berita fakta ataupun hoaks Disini peneliti memilih menggunakan IndoBERT sebagai ekstraksi kata ke numeric dan K- Means sebagai penentu klasterisasi berita tersebut.

a) Pernyataan Masalah

Bagaimana menerapkan model K-Means dan IndoBERT dapat digunakan untuk klasterisasi berita politik sebagai penentu berita hoaks atau fakta?

b) Tujuan Penelitian

Menerapkan model K-Means dan IndoBERT dalam membentuk representasi vektor dari berita politik berbahasa Indonesia guna sebagai penentu berita hoaks atau fakta.

c) Manfaat Penelitian

Mengembangkan metode klasterisasi berita hoaks dengan memanfaatkan kombinasi IndoBERT dan algoritma klasterisasi K-Means. Kemudian mengembangkan sistem informasi atau lembaga terkait seperti pemerintah, media, dan platform digital dalam menganalisis sebaran dan karakteristik berita hoaks yang beredar di Masyarakat. Meningkatkan transparansi dan kepercayaan terhadap informasi yang beredar di ruangpublik.

d) Batasan Masalah

Dataset yang diambil dari sumber media online berbahasa indonesia yang positif dan negatif yang tersedia secara publik melalui platform Kaggle tanpa memasukkan audio, judul, gambar atau video. Penelitian ini tidak membahas secara langsung proses deteksi kebenaran (verifikasi) berita hoaks, tetapi hanya fokus pada proses pengelompokan (clustering) berdasarkan kemiripan representasi teks berita.

2. Kajian Pustaka

a) Klasterisasi Berita

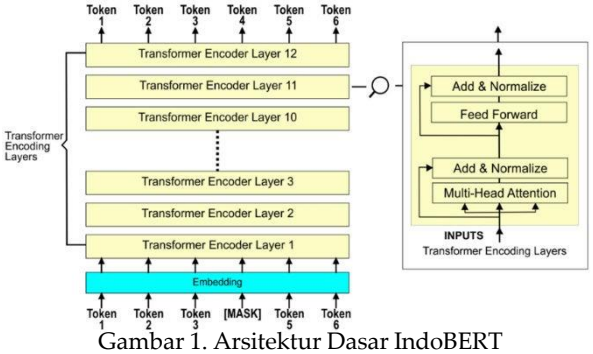
Klasterisasi merupakan teknik analisis data tanpa pengawasan yang penting, yang mencluster objek data ke dalam kelompok berdasarkan kesamaan. Klasterisasi telah dipelajari dan diterapkan di berbagai bidang, termasuk pengenalan pola, penambangan data, ilmu keputusan, dan statistik. Algoritma pengelompokan dapat diklasifikasikan terutama sebagai pendekatan pengelompokan hierarkis dan partisional[12]

Penelitian laian mengungkapkan bahwa klasterisasi berita merupakan topik penelitian yang sedang berkembang, yang mencakup dua bidang utama: NLP dan Machine Learning. Pengelompokan termasuk dalam kategori pembelajaran mesin tanpa pengawasan (unsupervised machine learning), yang lebih kompleks dalam hal implementasi dan evaluasi dibandingkan dengan varian yang diawasi (supervised). Terutama dalam domain dinamis seperti konten berita online, hal ini sangat sulit karena ketidakpastian label kluster atau jumlah kluster.

b) BERT dan IndoBERT

BERT (Bidirectional Encoder Representations from Transformers) merupakan model representasi bahasa berbasis arsitektur. BERT memanfaatkan pendekatan dua arah (bidirectional) dalam memproses teks, memungkinkan pemahaman konteks kata secara simultan dari kiri ke kanan dan sebaliknya. Hal ini berbeda dengan pendekatan left to right atau right to left pada model sebelumnya yang cenderung kehilangan konteks penuh. Untuk mencapai hal ini, BERT dilatih melalui dua mekanisme utama yaitu (MLM) dan (NSP).

[7] dalam penelitiannya mengatakan, beberapa tahun terakhir, model BERT telah digunakan untuk meneliti berbagai tugas klasifikasi dan klasterisasi dalam pemrosesan bahasa alami, seperti analisis sentimen, deteksi berita hoaks, dan tugas lainnya. Model BERT telah terbukti sangat efektif dalam analisis sentimen, seperti pada penelitian aplikasi Ruang Guru.

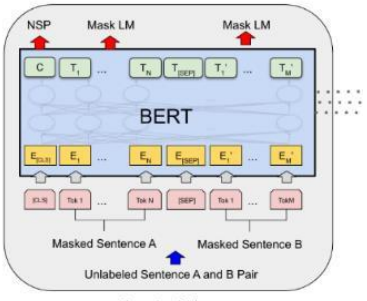


Gambar 1. Arsitektur Dasar IndoBERT

[Sumber: [13]]

Penulis membaca dari penelitian sebelumnya algoritma BERT (Bidirectional Encoder Representations from Transformers) merupakan algoritma deep learning yang berbasis NLP yang diluncurkan google pada tahun 2018 oleh Jacob Devlin dan rekan-rekannya dari Goolge Research berdasarkan arsitektur transformer yang yang dirancang untuk memahami konteks dari kata-kata sebuah teks secara dua arah (bidirectional). Sedang IndoBERT sendiri merupakan versi Bahasa Indonesia yang di kembangkan berdasarkan algoritma BERT dan mempunyai arsitektur yang sama, Dalam konteks penambangan teks, IndoBERT juga menunjukkan keunggulan dibandingkan RoBERTa, terutama dalam tugas pertanyaan dan jawaban menggunakan dataset terjemahan TyDi QA dan SQuAD. Namun, penggunaan dataset terjemahan dapat mengurangi akurasi dalam memahami konteks bahasa asli Indonesia, dan penelitian ini terbatas pada satu jenis aplikasi saja.

BERT dan IndoBERT menggunakan dua tahap utama dalam proses pembelajarannya: pre- training dan fine-tuning. Penulis mengambil satu Tahap pre-training yaitu dilakukan dengan data tanpa label menggunakan dua tugas prediktif, yaitu MLM dan NSP, maksudnya MLM yaitu melatih model untuk memahami konteks kata dalam kalimat, dengan memprediksi kata yang di-masking (disembunyikan) dari input teks. . Penulis mengambil satu Tahap pre-training yaitu dilakukan dengan data tanpa label menggunakan dua tugas prediktif, yaitu MLM dan NSP.



Pre-training

Gambar 2.2. Arsitektur dasar IndoBERT pre training

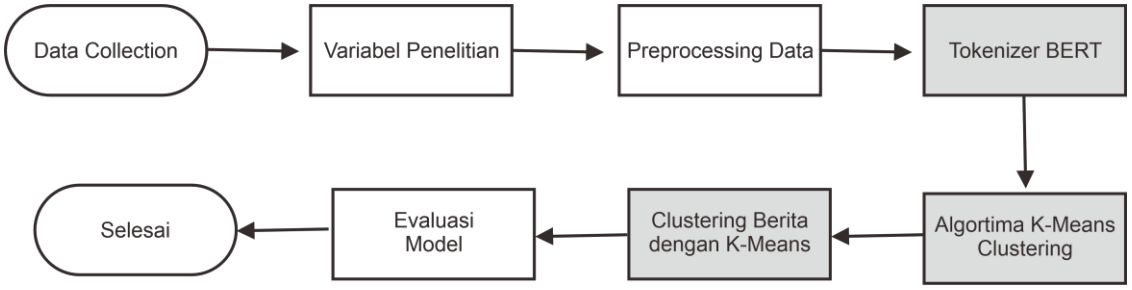
[Sumber: dari [13]]

c) K-Means Clustering

Clustering adalah teknik fundamental dalam data mining yang bertujuan untuk mengorganisir data ke dalam kelompokkelompok berdasarkan kesamaan karakteristik tertentu. Dalam konteks pengembangan sistem peringkasan teks otomatis, teknik clustering menjadi sangat penting untuk mengelompokkan kalimat-kalimat yang memiliki kemiripan makna atau topik, sehingga dapat menghasilkan ringkasan yang lebih terstruktur dan representatif. Sumber lain juga menyatakan salah satu algoritma clustering yang paling populer dan banyak digunakan karena kesederhanaannya dalam implementasi serta efektivitasnya dalam menangani dataset berukuran besar.

3. METODOLOGI PENELITIAN

Dalam penelitian ini, memiliki desain dengan penerapan klasterisasi berita politik bahasa Indonesia. Dengan mengimplementasikan metode pengelompokan ekstraktif. Desain sistem direpresentasikan dalam gambar di bawah ini.



Gambar 2.3 Desain Penelitian

Pada Gambar diatas terlihat desain sistem peringkasan teks dengan menggunakan pendekatan K- Means Clustering. Dalam sistem ini, dokumen artikel akan melalui tahapan Preprocessing untuk membersihkan teks, dilanjutkan dengan ekstraksi dengan IndoBERT pre-trained untuk merepresentasikan kalimat dalam bentuk vektor. Hasil pembobotan ini kemudian menjadi input untuk proses K-Means Clustering yang mengelompokkan kalimat serupa.

a. *Data Collection*

Tahap perencanaan merupakan langkah awal dalam pelaksanaan penelitian ini. Pada tahap ini, fokus diarahkan pada permasalahan meningkatnya penyebaran berita hoaks di media daring. Fenomena ini terjadi karena kurangnya pengetahuan masyarakat mengenai kebenaran informasi yang mereka bagikan. Oleh karena itu, diperlukan suatu solusi yang efektif guna menekan laju penyebaran berita hoaks tersebut yaitu dengan mengbuat deteksi untuk klasifikasi berita hoaks Bahasa Indonesia.

Langkah selanjutnya dalam penelitian adalah mengumpulkan dataset, Penelitian ini menggunakan dataset sekunder berupa kumpulan berita berbahasa Indonesia yang telah diberi label ke dalam dua kategori utama, yaitu berita faktual dan berita hoaks. Sumber data berasal dari tiga platform daring yang umum digunakan dalam penelitian data sains, yakni Kaggle, Mendeley, dan Huggingface. Ketiga sumber ini menyediakan dataset yang telah melalui proses penyaringan dan pengolahan serta pemberian label, sehingga data siap digunakan untuk keperluan analisis lebih lanjut. Proporsi data menunjukkan adanya ketidakseimbangan kelas, dengan 80% merupakan berita non-hoaks dan 20% merupakan berita hoaks.

Dataset ini dirancang untuk memfasilitasi penelitian, khususnya untuk mengembangkan model yang dapat membedakan antara pelaporan faktual dan cerita yang dibuat-buat dalam lingkup berita politik di Indonesia. Dataset ini mencakup metadata seperti sumber berita, title, dan narasi yang mengindikasikan apakah artikel tersebut faktual atau hoaks, sehingga menjadi sumber daya yang berharga bagi linguistik komputasi dan algoritma pendeteksian berita hoaks.

b. Variabel Penelitian

Bahan dan perlengkapan yang digunakan dalam melaksanakan penelitian ini adalah:

a. Personal Computer (PC) atau Laptop
b. Koneksi internet
c. Software pendukung seperti
   - Website: pencari dataset awal
   - Google Colab: sebagai media penulisan coding berbasis Python
   - Ms. Excel: digunakan sebagai pengolah data mentah dan penyimpaan data

c. Prepocessing Data

Tahapan pertama dalam proses preprocesing data adalah sebagai berikut:

a. Case folding

Tahapan selanjutnya adalah case folding yang akan mengubah seluruh kata yang menggunakan huruf kapital menjadi huruf kecil. Tujuan dari case folding adalah untuk menghindari perbedaan makna yang dapat terjadi ketika adanya huruf kapital suatu kata. Sebagai contoh, kata "Jakarta" dan "jakarta" seharusnya dianggap sama dalam pemrosesan teks, meskipun satu menggunakan huruf kapital di awal dan yang lainnya tidak. Dengan melakukan case folding, semua huruf dalam teks diubah menjadi huruf kecil, sehingga mengurangi kompleksitas analisis dan memastikan konsistensi dalam interpretasi kata-kata di seluruh teks.

Dengan demikian, case folding membantu meningkatkan konsistensi data dan mengurangi dimensi vektor representasi kata yang akan diproses lebih lanjut, seperti pada tahap tokenisasi dan embedding [14]

b. Noise removal

Langkah selanjutnya adalah noise removal. Noise removal adalah proses membersihkan teks dari elemen-elemen yang tidak relevan atau tidak memiliki nilai informasi dalam konteks analisis. Elemen ini meliputi tanda baca, simbol dan karakter, bisa dikatakan tahapan penting dalam pembersihan data teks untuk membuang unsur-unsur yang tidak memiliki kontribusi terhadap pemahaman konteks, seperti emoji, angka, tautan, tanda baca, dan karakter-karakter khusus.

Dengan menghilangkan noise, teks menjadi lebih bersih dan fokus pada konten yang relevan untuk analisis. Dengan tujuan adalah menyederhanakan teks sehingga lebih bersih dan sesuai untuk proses analisis lebih lanjut seperti tokenisasi dan pembentukan vektor teks[15]

c. Cleansing

Adalah menghapus elemen-elemen yang tidak relevan dalam analisis teks, seperti tautan atau URL. URL dianggap sebagai noise karena tidak memberikan kontribusi semantik terhadap isi berita, serta dapat mengganggu representasi vektor dalam pemrosesan lebih lanjut. Penghapusan URL dari teks berita agar informasi yang dianalisis hanya terdiri atas kata-kata bermakna secara linguistik [16]bisa dikatakan URL atau tautan yang terdapat dalam teks dihapus karena termasuk dalam kategori noise yang tidak memberikan makna semantik yang signifikan terhadap isi berita.

Tabel 1: Prepocessing Data
[Sumber:[17]]

No	Judul	Judul clean
1	[BENAR] KONDISI WILAYAHNYA DISEBUT TIDAK AMAN, BUPATI NDUGA PAPUA BERI KLARIFIKASI	benar kondisi wilayahnya disebut tidak aman bupati nduga papua beri klarifikasi
2	MENGETAHUI KANDUNGAN PASTA GIGI DARI KODE WARNA	mengetahui kandungan pasta gigi dari kode warna
3	JOKOWI JADI ORANG TERKAYA NOMOR 18 DI ASIA	jokowi jadi orang terkaya nomor 18 di asia
4	Tulisan Penolakan Omnibus Law oleh Dahlan Iskan	tulisan penolakan omnibus law oleh dahlan iskan
5	Foto Mobil Esemka Berbahan Dasar Kayu	foto mobil esemka berbahan dasar kayu
6	MODUS PERAMPOKAN ORANG TERGELETAK DI FATMAWATI	modus perampokan orang tergeletak di fatmawati
7	Info Lowongan Kerja di Puskesmas Wisma Indah, Bojonegoro Jawa Timur	info lowongan kerja di puskesmas wisma indah bojonegoro jawa timur
8	Bentrokan Pendukung Kubu 02 Dekat Kampus UKDW Tewaskan Dua Mahasiswa	bentrokan pendukung kubu 02 dekat kampus ukdw tewaskan dua mahasiswa
9	Maklumat KSAD Untuk Siaga Aksi 20 Oktober 2020	maklumat ksad untuk siaga aksi 20 oktober 2020
10	"jakarta menjadi daerah paling banyak terinfeksi virus covid-19 karena pilih gubernur"	jakarta menjadi daerah paling banyak terinfeksi virus covid 19 karena pilih gubernur

d. Tokenize BERT

Proses tokenisasi pada BERT bertujuan untuk mengubah teks mentah menjadi bentuk yang dapat diproses oleh model. Tokenisasi dilakukan dengan memecah kalimat menjadi bagian-bagian kecil yang disebut token. Pada BERT, tokenisasi menggunakan metode WordPiece,

yang secara otomatis memecah kata menjadi sub-kata atau morfem ketika kata tersebut tidak ditemukan dalam kosakata model. Proses ini diawali dengan menambahkan token khusus seperti [CLS] di awal dan [SEP] di akhir kalimat. Token yang dihasilkan kemudian dikonversi menjadi indeks numerik berdasarkan kamus BERT. Dengan demikian, model dapat memahami konteks dan makna dari kata-kata dalam kalimat.

Teks input berupa kalimat mentah akan melalui beberapa tahap pra pemrosesan sebelum dimasukkan ke dalam model BERT. Tahap pertama adalah tokenisasi dengan metode WordPiece untuk memecah kalimat menjadi unit-unit kecil (token). Kemudian, token khusus [CLS] ditambahkan di awal untuk menandai awal teks dan [SEP] di akhir untuk menandai batas akhir teks. Selanjutnya, token ini dikonversi menjadi indeks numerik (token ID) sesuai kamus BERT. Jika panjang kalimat kurang dari batas maksimum, padding dilakukan dengan menambahkan token [PAD] untuk menyamakan Panjang semua input. Terakhir, masking diterapkan untuk menandai posisi [PAD] agar diabaikan selama pemrosesan model.

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("indobenchmark/indobert-base-p1")
kalimat_list = [
    "di isi berita"]
encoded_batch = tokenizer(
    kalimat_list,
    padding=True,
    truncation=True,
    return_tensors="pt"
)
print("Token IDs:")
print(encoded_batch["input_ids"])
print("\nAttention Masks:")
print(encoded_batch["attention_mask"])
for i, kalimat in enumerate(kalimat_list):
    tokens = tokenizer.tokenize(kalimat)
    print(f"\nKalimat {i+1}: {kalimat}")
    print(f"Tokens: {tokens}")
```

Tabel 2: Tokenize BERT
[Sumber:[17]]

Langkah	Hasil
Teks Asli	benar kondisi wilayahnya disebut tidak aman bupati nduga papua beri klarifikasimengetahui kandungan pasta gigi dari kode warnajokowi jadi orang terkaya nomor 18 di asiatulisan penolakan omnibus law oleh dahlan iskanfoto mobil esemka berbahan dasar kayumodus perampokan orang tergeletak di fatmawatiinfo lowongan kerja di puskesmas wisma indah bojonegoro jawa timurbentrokan pendukung kubu 02 dekat kampus ukdw tewaskan dua mahasiswamaklumat ksad untuk siaga aksi 20 oktober 2020jakarta menjadi daerah paling banyak terinfeksi virus covid 19 karena pilih gubernurbenar kondisi wilayahnya disebut tidak aman bupati nduga papua beri klarifikasimengetahui kandungan pasta gigi dari kode warnajokowi jadi orang terkaya nomor 18 di asiatulisan penolakan omnibus law oleh dahlan iskanfoto mobil esemka berbahan dasar kayumodus perampokan orang tergeletak di fatmawatiinfo lowongan kerja di puskesmas wisma indah bojonegoro jawa timurbentrokan pendukung kubu 02 dekat kampus ukdw tewaskan dua mahasiswamaklumat ksad untuk siaga aksi 20 oktober 2020jakarta menjadi daerah paling banyak terinfeksi virus covid 19 karena pilih gubernur
Tokenisasi	['benar', 'kondisi', 'wilayahnya', 'disebut', 'tidak', 'aman', 'bupati', 'nd', '##uga', 'papua', 'beri', 'klarifikasi', '##meng', '##etahui', 'kandungan', 'pasta', 'gigi', 'dari', 'kode', 'warna', '##jok', '##owi',

	'jadi', 'orang', 'terkaya', 'nomor', '18', 'di', 'asia', '##tulis', '##an', 'penolakan', 'om', '##ni', '##bus', 'law', 'oleh', 'dahlan', 'iskan', '##foto', 'mobil', 'es', '##em', '##ka', 'berbahan', 'dasar', 'kayu', '##mod', '##us', 'perampokan', 'orang', 'tergeletak', 'di', 'fatmawati', '##info', 'lowongan', 'kerja', 'di', 'puskesmas', 'wisma', 'indah', 'bojonegoro', 'jawa', 'timur', '##ben', '##tro', '##kan', 'pendukung', 'kubu', '02', 'dekat', 'kampus', 'uk', '##d', '##w', 'tewas', '##kan', 'dua', 'mahasiswa', '##mak', '##lum', '##at', 'ks', '##ad', 'untuk', 'siaga', 'aksi', '20', 'oktober', '2020', '##jakarta', 'menjadi', 'daerah', 'paling', 'banyak', 'terinfeksi', 'virus', 'co', '##vid', '19', 'karena', 'pilih', 'gubernur', '##benar', 'kondisi', 'wilayahnya', 'disebut', 'tidak', 'aman', 'bupati', 'nd', '##uga', 'papua', 'beri', 'klarifikasi', '##meng', '##etahui', 'kandungan', 'pasta', 'gigi', 'dari', 'kode', 'warna', '##jok', '##owi', 'jadi', 'orang', 'terkaya', 'nomor', '18', 'di', 'asia', '##tulis', '##an', 'penolakan', 'om', '##ni', '##bus', 'law', 'oleh', 'dahlan', 'iskan', '##foto', 'mobil', 'es', '##em', '##ka', 'berbahan', 'dasar', 'kayu', '##mod', '##us', 'perampokan', 'orang', 'tergeletak', 'di', 'fatmawati', '##info', 'lowongan', 'kerja', 'di', 'puskesmas', 'wisma', 'indah', 'bojonegoro', 'jawa', 'timur', '##ben', '##tro', '##kan', 'pendukung', 'kubu', '02', 'dekat', 'kampus', 'uk', '##d', '##w', 'tewas', '##kan', 'dua', 'mahasiswa', '##mak', '##lum', '##at', 'ks', '##ad', 'untuk', 'siaga', 'aksi', '20', 'oktober', '2020', '##jakarta', 'menjadi', 'daerah', 'paling', 'banyak', 'terinfeksi', 'virus', 'co', '##vid', '19', 'karena', 'pilih', 'gubernur']
Konversi Token IDs	2, 839, 1186, 14935, 1082, 119, 1703, 3313, 6862, 2708, 4719, 5030, 19020, 5319, 973, 2674, 12456, 3168, 98, 2399, 1321, 10252, 4842, 472, 232, 17950, 1288, 1233, 26, 3173, 2233, 5, 11331, 3140, 784, 15554, 6127, 213, 12312, 24129, 6617, 895, 1660, 25, 3331, 8949, 1205, 2164, 7536, 50, 21657, 232, 29978, 26, 22499, 6575, 992, 494, 26, 7530, 15893, 1972, 14040, 1069, 1276, 1771, 11356, 32, 4627, 11376, 4857, 920, 4281, 1171, 30364, 30374, 6193, 32, 662, 1744, 8062, 7276, 15, 10435, 29, 90, 14461, 3668, 165, 2699, 18355, 6342, 234, 719, 711, 271, 9787, 3407, 2108, 7427, 426, 211, 959, 3256, 1935, 1186, 14935, 1082, 119, 1703, 3313, 6862, 2708, 4719, 5030, 19020, 5319, 973, 2674, 12456, 3168, 98, 2399, 1321, 10252, 4842, 472, 232, 17950, 1288, 1233, 26, 3173, 2233, 5, 11331, 3140, 784, 15554, 6127, 213, 12312, 24129, 6617, 895, 1660, 25, 3331, 8949, 1205, 2164, 7536, 50, 21657, 232, 29978, 26, 22499, 6575, 992, 494, 26, 7530, 15893, 1972, 14040, 1069, 1276, 1771, 11356, 32, 4627, 11376, 4857, 920, 4281, 1171, 30364, 30374, 6193, 32, 662, 1744, 8062, 7276, 15, 10435, 29, 90, 14461, 3668, 165, 2699, 18355, 6342, 234, 719, 711, 271, 9787, 3407, 2108, 7427, 426, 211, 959, 3256, 3]])
Attention Mask	1, 1]])

e. Algoritma K-Means Clustering

Algoritma K-Means dikenal merupakan metode klasterisasi yang paling sederhana untuk diimplementasikan dan dijalankan. Algoritma ini memiliki keunggulan dalam hal kemudahan adaptasi serta penggunaannya yang luas di berbagai bidang, terutama dalam praktik *data mining*. Sejarahnya, K-Means sendiri sebagai algoritma yang paling berpengaruh dalam pengembangan teknik klasterisasi non-hierarki, di mana algoritma ini bertujuan untuk membagi sekumpulan data ke dalam satu atau lebih kelompok (*cluster*) yang homogen [18]. Proses kerja algoritma K-Means dimulai dengan penentuan jumlah klaster atau nilai *k* yang didefinisikan di awal. Setelah itu, algoritma menghitung jarak antara setiap data (dalam konteks ini: representasi kata atau dokumen) dengan titik pusat klaster (*centroid*) yang telah ditentukan. Data kemudian diklasifikasikan ke dalam klaster yang memiliki jarak terdekat dengan centroid-nya, dan proses ini terus berulang hingga tercapai konvergensi atau kondisi stabil.

Secara umum, algoritma dasar K-Means Clustering adalah:

1. Tentukan banyaknya nilai K sebagai banyaknya cluster yang membentuk;
2. Tentukan secara acak jumlah nilai titik tengah (centroid) yang sama dengan jumlah cluster;
3. Menghitung jarak setiap masukan informasi ke setiap centroid dengan menggunakan rumus jarak Euclidean Distance sehingga ditemukan jarak terdekat dengan centroid disetiap informasi. Di bawah ini rumus persamaan jarak euclidean Distance:

$$d(xi_{\square}, \mu j_{\square}) = \sqrt{\sum_{\square}^{\square} (xi - \mu j_{\square})^2}$$

Dimana:

$xi_{\square}$ = data kriteria

$\mu j_{\square}$ = centroid pada cluster ke-j

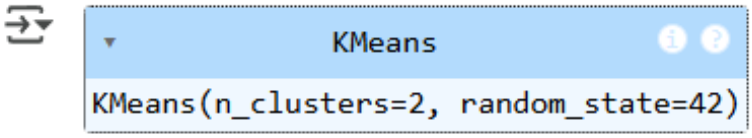
a) Inisiasi Centroid Awal

Pada tahap ini, sistem menginisialisasi titik pusat awal (centroid) untuk setiap cluster yang telah ditentukan. Jumlah k ditentukan melalui pencarian k optimal. Untuk percobaan ini, algoritma menggunakan K=2. Titik-titik pusat ini berfungsi sebagai acuan awal untuk mengelompokkan kalimat-kalimat berdasarkan kemiripan semantiknya.

```
from sklearn.cluster import KMeans


k = 2

kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(X)
```



Gambar 2.4 Inisiasi Centroid Awal

[Sumber: dari [17]]

Maksudnya yaitu algoritma K-Means Clustering menggunakan library scikit-learn, untuk mengelompokkan data vektor hasil embedding (misalnya dari IndoBERT) ke dalam 2 klaster yaitu hoaks dan Fakta. Berikut penjelasan tiap bagian:

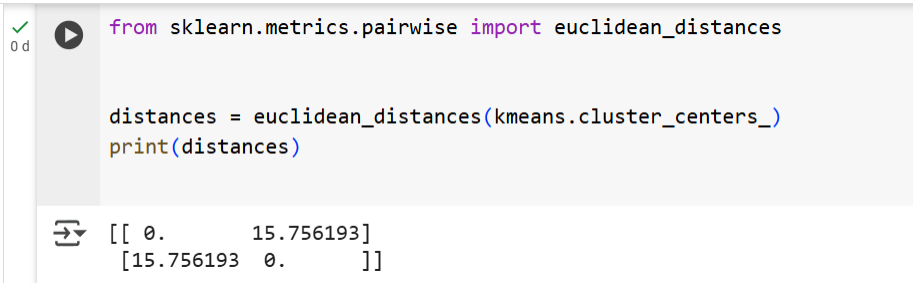
1. Membuat objek KMeans dengan parameter:
   - n_clusters=2: membagi data menjadi 2 kelompok.
   - init='k-means++': metode inisialisasi centroid yang lebih baik agar konvergensi lebih cepat dan hasil lebih stabil.
   - random_state=42: agar hasil klastering bisa direproduksi (hasil tidak berubah-ubah).
2. Melatih model KMeans pada data embeddings, yaitu array numpy hasil transformasi teks (biasanya vektor dari model BERT). Ini berarti kamu sedang mengelompokkan data embedding teks ke dalam 2 kelompok berdasarkan kesamaan vektor.

Hasil klaster bisa dilihat dengan kmeans.labels atau centroid-nya lewat kmeans.cluster centers.

b) Menghitung Jarak Centroid

Setelah centroid awal ditetapkan, sistem menghitung jarak setiap kalimat terhadap semua centroid menggunakan metrik jarak Euclidean seperti pada Persamaan 2.1. Jarak ini menjadi ukuran kemiripan antara kalimat dengan pusat cluster. Semakin kecil jarak antara kalimat dengan suatu centroid, semakin tinggi kemiripan semantiknya dengan cluster tersebut. Setiap kalimat kemudian ditetapkan ke cluster dengan jarak centroid terdekat yang dapat dilihat pada Tabel

Demikian pula, stemming, yang mengurangi kata menjadi bentuk dasarnya, tidak diperlukan karena IndoBERT dapat memahami bentuk kata yang berbeda dan maknanya yang beragam. Oleh karena itu, mempertahankan teks lengkap, termasuk stopwords dan bentuk kata, memungkinkan IndoBERT untuk memanfaatkan pembelajaran kontekstualnya yang mendalam, yang mengarah pada pemahaman teks yang lebih bernuansa dan akurat, yang sangat penting untuk tugas-tugas kompleks seperti klasifikasi.



Gambar 2.4 Hasil Jarak Cenroid

[Sumber: dari [17]]

c) Memperbarui Centroid

Tahap ini melibatkan perhitungan ulang pusat cluster (centroid) berdasarkan kalimat-kalimat yang telah dikelompokkan pada inisiasi sebelumnya menggunakan Persamaan 2.2. Centroid baru dihitung sebagai rata-rata vektor dari semua kalimat dalam cluster tersebut. Setelah centroid diperbarui, sistem kembali menghitung jarak dan mengelompokkan kalimat. Proses ini berulang hingga tidak ada perubahan yang signifikan pada posisi centroid (konvergen) atau jumlah iterasi maksimum tercapai

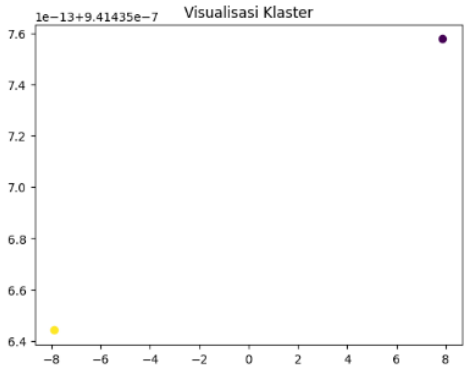
d) Hasil Cluster

Setelah algoritma konvergen, diperoleh hasil pengelompokan final dimana setiap kalimat telah ditetapkan ke salah satu dari 2 cluster. Setiap cluster berisi kalimat-kalimat dengan karakteristik semantik yang mirip berdasarkan representasi IndoBERT mereka. Hasil pengelompokan ini mencerminkan struktur semantik dari dokumen asli yang telah dipartisi ke dalam k kelompok berbeda.

Setelah proses K-Means selesai dilakukan, hasil clustering dapat diinterpretasikan melalui beberapa hal berikut:

1. Label Klaster (kmeans.labels) Setiap data akan diberi label berdasarkan klaster tempat ia tergabung (misalnya 0 atau 1 untuk dua klaster). Ini menunjukkan bahwa data tersebut paling dekat (jarak Euclidean terkecil) ke centroid klaster tersebut.
2. Titik Pusat (Centroid) (kmeans.cluster_centers) Menunjukkan posisi rata-rata (mean) dari semua data dalam setiap klaster. Ini dapat digunakan untuk memahami karakteristik umum dari tiap kelompok.
3. Visualisasi

   Untuk memudahkan interpretasi, hasil klaster dapat divisualisasikan (jika data berdimensi rendah seperti 2D atau 3D). Jika Anda menggunakan vektor embedding seperti dari IndoBERT, Anda bisa melakukan reduksi dimensi terlebih dahulu (misalnya dengan PCA atau TSNE) sebelum divisualisasi.



Gambar 2.5 Hasil Jarak Cenroid

[Sumber: dari [17]]

Hasil klaster ini bisa digunakan untuk analisis lanjutan seperti deteksi topik berita, penyaringan data hoaks, atau segmentasi pengguna.

4. Hasil Dan Pembahasan

a) Pengukuran Jarak Dengan Consine Similarity

Cosine Similarity adalah metode untuk mengukur tingkat kemiripan antara dua vektor dalam ruang berdimensi tinggi dengan menghitung cosinus sudut antara keduanya. Nilai kemiripan ini berkisar antara -1 hingga 1, tetapi dalam konteks representasi teks (yang hasil embedding-nya bernilai positif), nilai cosine similarity umumnya berada dalam rentang 0 hingga 1, dengan:

- Nilai 1 menunjukkan bahwa dua vektor memiliki arah yang sama (sangat mirip).
- Nilai 0 menunjukkan bahwa dua vektor tegak lurus (tidak ada kesamaan).
- Nilai mendekati 0 atau negatif menunjukkan ketidaksamaan.

Rumus cosine similarity

$$(\mathrm{dj,q}) = \frac{\sum_{i=l}^{t}(w_{ij}.w_{iq})}{\sqrt{\sum_{i=l}^{t} w_{ij}^2 . \sum_{i=l}^{t} w_{iq}^2}}$$

b) Hasil Representasi Teks dengan IndoBERT

Hasil representasi teks dari IndoBERT selanjutnya digunakan sebagai input dalam algoritma K-Means, di mana tiap vektor embedding mewakili fitur numerik dari masing-masing teks berita untuk proses pengelompokan secara tidak terawasi. Setelah data dibersihkan, representasi vektor dibuat menggunakan IndoBERT dengan memanfaatkan token [CLS] sebagai representasi dokumen. Vektor yang dihasilkan memiliki dimensi 768, sesuai dengan arsitektur IndoBERT.

Gambar 2.6 Hasil Representasi Text ke IndoBERT
[Sumber: dari [17]]

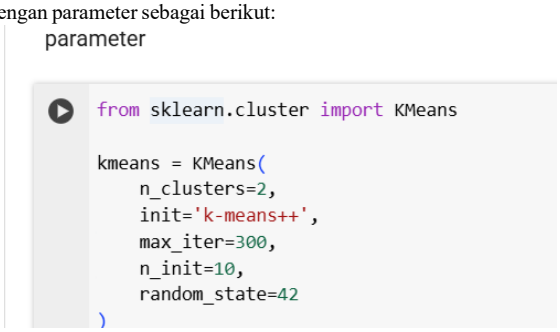
Penjelasan:
1. Tokenizer dan Model: IndoBERT digunakan untuk mengubah teks ke dalam bentuk vektor numerik (embedding).
2. Tokenisasi: Teks diubah menjadi format input untuk BERT.
3. Embedding: Mengambil representasi dari token [CLS] untuk tiap teks.
4. K-Means: Vektor embedding dikelompokkan menjadi 2 cluster menggunakan K-Means.
5. Output: Menampilkan label klaster dari setiap teks.

c) Klasterisasi Berita dengan K-Means

Dalam tahap klasterisasi ini, vektor representasi dari teks yang dihasilkan oleh IndoBERT akan dikelompokkan menggunakan algoritma K-Means Clustering, dengan tujuan memisahkan data ke dalam dua kategori utama, yaitu:

- Cluster 0: Diinterpretasikan sebagai berita Fakta
- Cluster 1: Diinterpretasikan sebagai berita Hoaks

Model KMeans diinisialisasi dengan parameter sebagai berikut:



Gambar 2.7 Klasterisasi Berita dengan K-Means
[Sumber: dari [17]]

Penjelasan parameter:

1. n_clusters=2: Menetapkan jumlah klaster ke 2 (Fakta dan Hoaks) sesuai dengan tujuan penelitian
2. init='k-means++': Inisialisasi centroid awal yang lebih cerdas, untuk mempercepat konvergensi dan meningkatkan akurasi
3. max_iter=300: Batas maksimum iterasi untuk memastikan proses konvergensi optimal.
4. n_init=10: Algoritma akan dijalankan 10 kali dengan inisialisasi berbeda, dan hasil terbaik dipilih.
5. random_state=42: Agar hasil eksperimen dapat direproduksi.

```
from transformers import AutoTokenizer, AutoModel
import torch

tokenizer = AutoTokenizer.from_pretrained("indobenchmark/indobert-base-p1")
model = AutoModel.from_pretrained("indobenchmark/indobert-base-p1")

texts = [
    "Presiden Jokowi meresmikan bendungan Tukul di Pacitan sebagai upaya memperkuat ketahanan air.",
    "Vaksin COVID-19 mengandung chip pelacak untuk mengontrol pikiran manusia."
]

inputs = tokenizer(texts, padding=True, truncation=True, return_tensors="pt")
with torch.no_grad():
    outputs = model(**inputs)

vectors = outputs.last_hidden_state[:, 1, :]
```

```
Asking to truncate to max_length but no maximum length is provided and the model has no predefined maximum length. Default to no truncation.
```

```
[94] from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=2, random_state=42)
kmeans.fit(vectors.numpy())


print(kmeans.labels_)  # Output misalnya: [0 1]
```

```
[0 1]
```

Gambar 2.8 Hasil Parameter Berita dengan K Means [Sumber: dari [17]]

d) Evaluasi Hasil Klasterisasi

Untuk menilai kualitas hasil klasterisasi, digunakan metrik Silhouette Score. Metrik ini mengukur seberapa baik suatu objek cocok dengan klaster yang menjadi anggotanya dibandingkan dengan klaster lainnya. Nilai Silhouette Score berkisar antara -1 hingga 1. Semakin mendekati nilai 1, maka kualitas pemisahan antar klaster semakin baik, karena objek lebih dekat dengan klasternya sendiri daripada dengan klaster lain. Sebaliknya, nilai yang mendekati 0 menunjukkan adanya tumpang tindih antar klaster, dan nilai negatif menunjukkan bahwa objek lebih mirip dengan klaster lain dibandingkan dengan klaster tempat ia ditempatkan.

```
from sklearn.metrics import silhouette_score

score = silhouette_score(X,
kmeans.labels_) print(f"Silhouette
Score: {score:.4f}")
```

Silhouette Score: 0.2105

Evaluasi Hasil Klasterisasi Menggunakan Silhouette Score

a. Salah satu metrik yang digunakan untuk mengevaluasi kualitas hasil klasterisasi adalah Silhouette Score. Metrik ini mengukur seberapa baik suatu data cocok dengan klasternya sendiri dibandingkan dengan klaster lainnya. Nilai Silhouette Score berada pada rentang –1 hingga +1, dengan interpretasi sebagai berikut:
b. Nilai mendekati +1 menunjukkan bahwa data sangat cocok dengan klasternya dan tidak cocok dengan klaster lain (klasterisasi sangat baik).
c. Nilai sekitar 0 menunjukkan bahwa data berada di batas antara dua klaster (klasterisasi kurang jelas).
d. Nilai mendekati –1 menunjukkan bahwa data lebih cocok berada di klaster lain (klasterisasi buruk).

Dalam konteks klasterisasi berita:

a. Silhouette Score tinggi (misalnya > 0.5) menandakan bahwa berita-berita dalam satu klaster memiliki kesamaan topik atau konteks yang kuat, dan berbeda signifikan dengan berita di klaster lain. Ini berarti sistem klasterisasi mampu memisahkan berita faktual dan hoaks dengan cukup akurat.
b. Silhouette Score rendah (misalnya < 0.2) menunjukkan bahwa sistem kesulitan dalam membedakan antar topik, sehingga kemungkinan terjadi tumpang tindih antar klaster, yang mengurangi kegunaan hasil klasterisasi bagi pengguna akhir.
c. Jika nilai negatif, itu menunjukkan kesalahan pengelompokan, di mana berita sering kali dimasukkan ke dalam klaster yang tidak sesuai dengan isinya. Dalam praktiknya, ini menunjukkan bahwa model atau parameter perlu diperbaiki.

Silhouette Score sebesar 0,21 menunjukkan bahwa sistem klasterisasi IndoBERT + K-Means ini belum mencapai klasterisasi sempurna, namun sudah cukup efektif dalam membedakan dua kelompok utama berita, yaitu hoaks dan fakta. Ini menjadikan pendekatan tersebut layak untuk digunakan dalam sistem klasifikasi awal berita tanpa label (unsupervised), dengan potensi besar untuk ditingkatkan lebih lanjut menggunakan data dan pendekatan yang lebih kompleks.

5. Kesimpulan

Penelitian ini bertujuan untuk mengelompokkan berita berbahasa Indonesia ke dalam dua kategori utama, yaitu berita hoaks dan fakta, menggunakan metode K-Means Clustering dan representasi teks dari IndoBERT. Berdasarkan hasil yang diperoleh, IndoBERT mampu menghasilkan representasi vektor (embedding) dari teks berita dengan baik, karena model ini telah dilatih secara khusus untuk memahami konteks semantik dalam Bahasa Indonesia. Proses pra-pemrosesan teks, yang meliputi pembersihan data, case folding, dan penggabungan judul serta isi berita, berhasil menghasilkan input yang bersih dan sesuai untuk model representasi IndoBERT. Metode K-Means Clustering berhasil mengelompokkan berita ke dalam dua klaster. Setelah dilakukan analisis terhadap isi klaster, dapat diinterpretasikan bahwa salah satu klaster didominasi oleh berita hoaks, sedangkan klaster lainnya berisi berita fakta. Evaluasi menggunakan Silhouette Score sebesar 0.21 menunjukkan bahwa hasil klasterisasi berada dalam kategori kurang baik, yang berarti pemisahan antar berita dalam masing-masing klaster cukup jelas secara semantik. Dengan pendekatan unsupervised learning, sistem ini mampu mengelompokkan berita tanpa memerlukan label manual, dan tetap menghasilkan pembagian yang sesuai dengan sifat berita hoaks dan fakta berdasarkan analisis isi.

6. Daftar Pustaka

[1] R. Yunanto, A. P. Purfini, and A. Prabuwisesa, "Survei Literatur: Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning," *J. Manaj. Inform.*, vol. 11, no. 2, pp. 118–130, 2021, doi: 10.34010/jamika.v11i2.5362.

[2] 2019) (EMCHA, WIDYAWAN, & ADJI, "Klasterisasi Berita Bahasa Indonesia Dengan Menggunakan K-Means Dan Word Embedding," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 3, pp. 641–652, 2019, doi: 10.25126/jtiik.20231026468.

[3] D. Ramadanti *et al.*, "Perspektif Masyarakat Terhadap Humas Dalam Penggunaan Media Sosial," *Al-Zayn J. Ilmu Sos. Huk.*, vol. 3, pp. 165– 171, 2025, doi: 10.61104/alz.v3i2.919.

[4] 1-9. *et al.*, "Scholar (3)," *Annals of Tourism Research*, vol. 3, no. 1. pp. 1–2, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0160738315000444

[5] A. L. Basuki *et al.*, "ANALISIS SENTIMEN ULASAN APLIKASI AJAIB KRIPTO," vol. 9, no. 4, pp. 1–9, 2025.

[6] A. Kunaefi, Z. Abidin, and R. Kusumawati, "Klasifikasi berita hoaks bahasa indonesia menggunakan indobert," vol. 10, no. 2, pp. 1706–1714, 2025.

[7] C. J. L. Tobing, I. G. N. L. Wijayakusuma, and L. P. Ida, "Perbandingan Kinerja IndoBERT dan MBERT untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia," vol. 14, no. 1, pp. 114–123, 2025.

[8] I. M. S. Bimantara and I. M. Widiartha, "Optimization of K-Means Clustering Using Particle Swarm Optimization Algorithm for Grouping Traveler Reviews Data on Tripadvisor Sites," *J. Ilm. Kursor*, vol. 12, no. 1, pp. 1–10, 2023.

[9] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.

[10] R. R. Putra and C. Wadisman, "Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K Means," *Intecoms*, vol. 1, no. 1, pp. 72–77, 2018, doi: 10.31539/intecoms.v1i1.141.

[11] F. Kamila, E. Prasetyo, and W. Roessali, "ANALISIS SIKAP KONSUMEN PADA PEMBELIAN BERAS DI KOTA SALATIGA," *Agrisocionomics J. Sos. Ekon. Pertan.*, vol. 3, p. 10, 2019, doi: 10.14710/agrisocionomics.v3i1.2980.

[12] A. Onan, "A K-medoids based clustering scheme with an application to document clustering," in *2017 international conference on computer science and engineering (UBMK)*, IEEE, 2017, pp. 354–359.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[14] A. Ripa'i, F. Santoso, and F. Lazim, *Deteksi Berita Hoax dengan Perbandingan Website Menggunakan Pendekatan Deep Learning Algoritma BERT*, vol. 8, no. 3. 2024. doi: 10.33379/gtech.v8i3.4541.

[15] H. Ulfatriyani, H. A. Nugroho, and I. Soesanti, "Implementing Term Frequency-Inverse Term Frequency at Tweets in Indonesian Fraud Crime Cases," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 2020, pp. 185–190. doi: 10.1109/ICOIACT50329.2020.9331996.

[16] I. H. Witten, E. Frank, and M. A. Hall, "Chapter 4 - Algorithms: The Basic Methods," in *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*, Third Edit., I. H. Witten, E. Frank, and M. A. Hall, Eds., in The Morgan Kaufmann Series in Data Management Systems. , Boston: Morgan Kaufmann, 2011, pp. 85–145. doi: https://doi.org/10.1016/B978-0-12-374856-0.00004-3.

[17] G. Colab, "Google Colab IndoBERT." 2024.

[18] H. Sulastri and A. I. Gufroni, "Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017, doi: 10.25077/teknosi.v3i2.2017.299-30