

ANALISIS SEBARAN TOPIK ARTIKEL ILMIAH
MENGUNAKAN INDOBERT DAN K-MEANS

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika



Oleh :

SITI MARITZA AQILA
NIM: 09021282126048

Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2024

LEMBAR PENGESAHAN SKRIPSI

**ANALISIS SEBARAN TOPIK ARTIKEL ILMIAH MENGGUNAKAN
INDOBERT DAN K-MEANS**

Oleh :

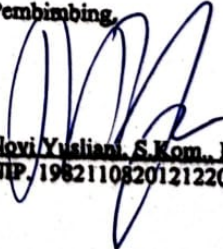
Siti Maritza Aqila
NIM: 09021282126048

Mengetahui,



Indrasya, 31 Desember 2024

Pembimbing,



Novi Yuliana, S.Kom., M.T.
NIP. 198211082012122001

TANDA LULUS UJIAN KOMPREHENSIF

Pada hari Selasa tanggal 31 Desember 2024 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Siti Maritza Aqila
Nim : 09021282126048
Judul : Analisis Sebaran Topik Artikel Ilmiah Menggunakan IndoBERT dan K-Means

Dan dinyatakan LULUS.

1. Ketua Penguji

Rizki Kurniati, S.Kom., M.T.
NIP. 199107122019032016

2. Penguji I

M. Qurhanul Rizqie, M.T., Ph.D.
NIP. 198712032022031006

3. Pembimbing

Novi Yustiani, S.Kom., M.T.
NIP. 198211082012122001

Mengotahui,
Ketua Jurusan Teknik Informatika



Madhuzwan Setris, Ph.D.
NIP. 196004182020121001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Siti Maritza Aqila
NIM : 09021282126048
Program Studi : Teknik Informatika
Judul Skripsi : Analisis Sebaran Topik Artikel Ilmiah Menggunakan
IndoBERT dan K-Means

Hasil Pengecekan *Software iThenticate/Turnitin*: 5%

Menyatakan bahwa laporan penelitian saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam penelitian ini maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.



Palembang, 6 Januari 2024



Siti Maritza Aqila

NIM. 09021282126048

MOTTO DAN PERSEMBAHAN

“Setiap halaman yang ditulis adalah bukti perjuangan, doa, dan dukungan tulus dari orang-orang terdekat.”

Saya persembahkan karya tulis ini kepada:

- Orangtua
- Keluarga Besar
- Sahabat
- Teman Seperjuangan
- Fakultas Ilmu Komputer
- Universitas Sriwijaya

Abstract

The significant increase in the number of scientific articles published along with advances in science and technology poses challenges in managing and organizing articles to support the efficiency of literature analysis. This research aims to develop a scientific article topic distribution analysis system by utilizing the IndoBERT model and the K-Means algorithm. The dataset used comes from 12 nationally accredited scientific journals on the Science and Technology Index (SINTA). Clustering evaluation was conducted using silhouette score, with the highest value of 0.932656 in journals with EISSN 25799258. This clustering model is then applied to predict new articles based on the relevance of their journal topics, with inscope or outscope categories determined based on the threshold value.

Keywords: *topic distribution analysis of scientific articles, IndoBERT, K-Means, silhouette score, article prediction.*

Abstrak

Peningkatan jumlah artikel ilmiah yang diterbitkan secara signifikan seiring dengan kemajuan ilmu pengetahuan dan teknologi menimbulkan tantangan dalam pengelolaan dan pengorganisasian artikel untuk mendukung efisiensi analisis literatur. Penelitian ini bertujuan untuk mengembangkan sistem analisis sebaran topik artikel ilmiah dengan memanfaatkan model IndoBERT dan algoritma K-Means. Dataset yang digunakan berasal dari 12 jurnal ilmiah terakreditasi nasional pada *Science and Technology Index (SINTA)*. Evaluasi klusterisasi dilakukan menggunakan *silhouette score*, dengan nilai tertinggi sebesar 0,932656 pada jurnal dengan EISSN 25799258. Model klusterisasi ini selanjutnya diterapkan untuk memprediksi artikel baru berdasarkan relevansi topik jurnalnya, dengan kategori *inscope* atau *outscope* yang ditentukan berdasarkan nilai ambang batas (*threshold*).

Kata kunci: analisis sebaran topik artikel ilmiah, IndoBERT, K-Means, *silhouette score*, prediksi artikel.

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan kepada Allah SWT atas limpahan rahmat dan karunia Nya sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Penulisan skripsi ini dilakukan sebagai salah satu syarat untuk menyelesaikan pendidikan program Strata-1 di Fakultas Ilmu Komputer Universitas Sriwijaya. Dalam proses penyelesaian skripsi ini, penulis memperoleh berbagai bantuan, bimbingan, serta dukungan dari berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih kepada:

1. Allah SWT atas rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi ini dengan baik.
2. Kedua orang tua dan keluarga besar yang telah mendoakan, memberi semangat, motivasi dan nasihat dalam menyelesaikan skripsi ini.
3. Bapak Hadipurnawan Satria, S.Kom. M.Sc, Ph.D. selaku Ketua Jurusan Teknik Informatika Universitas Sriwijaya.
4. Bapak Rifkie Primartha, S.T., M.T. selaku Dosen Pembimbing Akademik yang telah memberikan bantuan dan arahan kepada penulis selama perkuliahan.
5. Ibu Novi Yusliani, S.Kom., M.T. selaku Dosen Pembimbing yang telah membimbing serta memberikan arahan kepada penulis selama proses pengerjaan skripsi.
6. Seluruh dosen program studi serta admin Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Seluruh Staf Administrasi dan Pegawai Fakultas Ilmu Komputer yang telah membantu dalam urusan administrasi tugas akhir penulis.
8. Seluruh sahabat dan teman-teman yang telah membantu, memberikan saran, motivasi, dan semangat kepada penulis khususnya Rachma, Valyssa, Yolen, Oca, Ori, dan Meutia.
9. Pihak-pihak lain yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih banyak sekali kekurangan terhadap penyusunan tugas akhir ini. Oleh karena itu, penulis mengharapkan saran dan kritik yang membangun guna kemajuan penelitian kedepannya. Semoga tugas akhir ini dapat bermanfaat. Terima kasih.

Palembang, 24 Desember 2024

Penulis,



Siti Maritza Aqila

DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI	Error! Bookmark not defined.
TANDA LULUS UJIAN KOMPREHENSIF	Error! Bookmark not defined.
HALAMAN PERNYATAAN.....	Error! Bookmark not defined.
MOTTO DAN PERSEMBAHAN.....	v
Abstract.....	vi
Abstrak.....	vii
KATA PENGANTAR	viii
DAFTAR ISI	x
DAFTAR TABEL	xiv
DAFTAR GAMBAR.....	xvii
BAB I.....	1
PENDAHULUAN	I-1
1.1 Pendahuluan.....	I-1
1.2 Latar Belakang.....	I-1
1.3 Rumusan Masalah.....	I-3
1.4 Tujuan Penelitian	I-3
1.5 Manfaat Penelitian	I-3
1.6 Batasan Masalah	I-4
1.7 Sistematika Penulisan	I-4
1.8 Kesimpulan.....	I-6
BAB II	II-Error! Bookmark not defined.
KAJIAN LITERATUR.....	II-Error! Bookmark not defined.
2.1 Pendahuluan.....	II-Error! Bookmark not defined.

2.2	Landasan Teori	II-Error! Bookmark not defined.
2.2.1	Klasterisasi Teks	II-Error! Bookmark not defined.
2.2.2	K-Means Clustering	II-Error! Bookmark not defined.
2.2.3	BERT	II-Error! Bookmark not defined.
2.2.3.1	IndoBERT	II-Error! Bookmark not defined.
2.2.4	Principal Component Analysis	II-Error! Bookmark not defined.
2.2.5	Silhouette Score	II-Error! Bookmark not defined.
2.2.6	Rational Unified Process	II-Error! Bookmark not defined.
2.3	Penelitian Lain yang Relevan	II-Error! Bookmark not defined.
2.4	Kesimpulan.....	II-Error! Bookmark not defined.
BAB III		III-Error! Bookmark not defined.
METODE PENELITIAN		III-Error! Bookmark not defined.
3.1	Pendahuluan.....	III-Error! Bookmark not defined.
3.2	Pengumpulan Data.....	III-Error! Bookmark not defined.
3.2.1	Jenis dan Sumber Data.....	III-Error! Bookmark not defined.
3.2.2	Metode Pengumpulan Data.....	III-Error! Bookmark not defined.
3.3	Tahapan Penelitian.....	III-Error! Bookmark not defined.
3.3.1	Mengumpulkan Data	III-Error! Bookmark not defined.
3.3.2	Menentukan Kerangka Kerja Penelitian	III-Error! Bookmark not defined.
		defined.
3.3.3	Menentukan Kriteria Pengujian	III-Error! Bookmark not defined.
3.3.4	Menentukan Format Data Pengujian	III-Error! Bookmark not defined.
		defined.
3.3.5	Menentukan Alat Bantu Penelitian	III-Error! Bookmark not defined.
3.3.6	Melakukan Pengujian Penelitian ..	III-Error! Bookmark not defined.
3.3.7	Melakukan Analisis dan Menarik Kesimpulan	III-Error! Bookmark not defined.
		not defined.
3.4	Metode Pengembangan Perangkat Lunak	III-Error! Bookmark not defined.
		defined.

3.4.2	Fase Elaborasi.....	III-Error! Bookmark not defined.
3.4.3	Fase Konstruksi	III-Error! Bookmark not defined.
3.4.4	Fase Transisi	III-Error! Bookmark not defined.
3.5	Kesimpulan.....	III-Error! Bookmark not defined.
BAB IV		IV-Error! Bookmark not defined.
PENGEMBANGAN PERANGKAT LUNAK		IV-Error! Bookmark not defined.
4.1	Pendahuluan.....	IV-Error! Bookmark not defined.
4.2	Fase Insepsi.....	IV-Error! Bookmark not defined.
4.2.2	Kebutuhan Sistem.....	IV-Error! Bookmark not defined.
4.2.3	Analisis dan Desain	IV-Error! Bookmark not defined.
4.2.3.1	Analisis Kebutuhan Perangkat Lunak.....	IV-Error! Bookmark not defined.
4.2.3.2	Analisis Data.....	IV-Error! Bookmark not defined.
4.2.3.3	Analisis Text Pre-Processing.....	IV-Error! Bookmark not defined.
4.2.3.4	Analisis Model Bahasa IndoBERT.....	IV-Error! Bookmark not defined.
4.2.3.5	Analisis Principal Component Analysis (PCA).....	IV-Error! Bookmark not defined.
4.2.3.6	Analisis K-Means	IV-Error! Bookmark not defined.
4.2.3.7	Analisis Silhouette Score	IV-Error! Bookmark not defined.
4.2.3.8	Desain Perangkat Lunak	IV-Error! Bookmark not defined.
4.3	Fase Elaborasi.....	IV-Error! Bookmark not defined.
4.3.1	Pemodelan Bisnis.....	IV-Error! Bookmark not defined.
4.3.1.1	Perancangan Data	IV-Error! Bookmark not defined.
4.3.1.2	Desain Antarmuka	IV-Error! Bookmark not defined.
4.3.2	Kebutuhan Sistem.....	IV-Error! Bookmark not defined.
4.3.3	Analisis dan Perancangan	IV-Error! Bookmark not defined.
4.3.3.1	Diagram Activity	IV-Error! Bookmark not defined.
4.3.3.2	Diagram Sequence	IV-Error! Bookmark not defined.

4.4	Fase Konstruksi	IV-Error! Bookmark not defined.
4.4.1	Kebutuhan Sistem	IV-Error! Bookmark not defined.
4.4.2	Implementasi.....	IV-Error! Bookmark not defined.
4.4.2.1	Implementasi Kelas.....	IV-Error! Bookmark not defined.
4.4.2.2	Implementasi Antarmuka.....	IV-Error! Bookmark not defined.
4.5	Fase Transisi	IV-Error! Bookmark not defined.
4.5.1	Pemodelan Bisnis.....	IV-Error! Bookmark not defined.
4.5.2	Rencana Pengujian.....	IV-Error! Bookmark not defined.
4.5.3	Implementasi.....	IV-Error! Bookmark not defined.
4.6	Kesimpulan.....	IV-Error! Bookmark not defined.
BAB V		V-Error! Bookmark not defined.
HASIL DAN PEMBAHASAN		V-Error! Bookmark not defined.
5.1	Pendahuluan.....	V-Error! Bookmark not defined.
5.2	Hasil Penelitian	V-Error! Bookmark not defined.
5.2.1	Konfigurasi Pengujian	V-Error! Bookmark not defined.
5.2.2	Data Hasil Konfigurasi	V-Error! Bookmark not defined.
5.3	Analisis Hasil Pengujian.....	V-Error! Bookmark not defined.
5.4	Analisis Hasil Klasterisasi	V-Error! Bookmark not defined.
5.5	Analisis Hasil Prediksi Klaster Artikel Baru	V-Error! Bookmark not defined.
	defined.	
5.6	Kesimpulan	V-Error! Bookmark not defined.
BAB VI.....		V1-Error! Bookmark not defined.
KESIMPULAN DAN SARAN		V1-Error! Bookmark not defined.
6.1	Pendahuluan.....	VI-Error! Bookmark not defined.
6.2	Kesimpulan	VI-Error! Bookmark not defined.
6.3	Saran	VI-Error! Bookmark not defined.

DAFTAR PUSTAKA.....xvii

DAFTAR TABEL

Tabel III-1. Contoh Data Artikel Ilmiah.....	III-2
Tabel III-2. Hasil Pengujian.....	III-10
Tabel III-3. Alat Bantu Penelitian.....	III-10
Tabel III-4. Hasil Analisis Pengujian	III-11
Tabel IV-1. Kebutuhan Fungsional	IV-Error! Bookmark not defined.
Tabel IV-2. Kebutuhan Non-Fungsional	IV-Error! Bookmark not defined.
Tabel IV- 3. Contoh Data Judul dan Abstrak ..	IV-Error! Bookmark not defined.
Tabel IV-4. Hasil Cleaning dan Case Folding	IV-Error! Bookmark not defined.
Tabel IV-5. Hasil Language Detection.....	IV-Error! Bookmark not defined.
Tabel IV-6. Hasil Tokenizing dan Embedding dengan IndoBERT.....	IV-Error! Bookmark not defined.
Tabel IV-7. Contoh Hasil Principal Component Analysis (PCA).....	IV-Error! Bookmark not defined.
Tabel IV-8. Hasil Evaluasi Silhouette Score ..	IV-Error! Bookmark not defined.
Tabel IV-9. Tabel Definisi Aktor	IV-Error! Bookmark not defined.
Tabel IV-10. Tabel Definisi Use Case Pengelompokan Artikel Ilmiah Menggunakan IndoBERT dan K-Means .	IV-Error! Bookmark not defined.
Tabel IV-11. Tabel Definisi Use Case Prediksi Klaster Pengelompokan Artikel Ilmiah Menggunakan IndoBERT dan K-Means.....	IV-Error! Bookmark not defined.
Tabel IV-12. Skenario Use Case Melakukan Klasterisasi.....	IV-Error! Bookmark not defined.
Tabel IV-13. Skenario Use Case Evaluasi Hasil	IV-Error! Bookmark not defined.
Tabel IV-14. Skenario Use Case Memasukkan Data	IV-Error! Bookmark not defined.
Tabel IV-15. Skenario Use Case Melakukan Prediksi Klaster.....	IV-Error! Bookmark not defined.
Tabel IV-16. Implementasi Kelas.....	IV-Error! Bookmark not defined.

Tabel IV-17. Rencana Pengujian Use Case Melakukan KlasterisasiIV-**Error! Bookmark not defined.**

Tabel IV-18. Rencana Pengujian Use Case Evaluasi Hasil...IV-**Error! Bookmark not defined.**

Tabel IV-19. Rencana Pengujian Use Case Memasukkan Data.....IV-**Error! Bookmark not defined.**

Tabel IV-20. Rencana Pengujian Use Case Prediksi Klaster IV-**Error! Bookmark not defined.**

Tabel IV-21. Pengujian Use Case Melakukan Klasterisasi ...IV-**Error! Bookmark not defined.**

Tabel IV- 22. Pengujian Use Case Proses Evaluasi HasilIV-**Error! Bookmark not defined.**

Tabel IV-23. Pengujian Use Case Proses Memasukkan DataIV-**Error! Bookmark not defined.**

Tabel IV-24. Pengujian Use Case Proses Prediksi Klaster....IV-**Error! Bookmark not defined.**

Tabel V-1. Tabel Nilai Evaluasi V-**Error! Bookmark not defined.**

Tabel V-2. Hasil Prediksi Klaster EISSN 23028556 (Jurnal Akuntansi) .. V-**Error! Bookmark not defined.**

Tabel V-3. Hasil Prediksi Klaster EISSN 23553596 (Jurnal Kesehatan Masyarakat) V-**Error! Bookmark not defined.**

Tabel V-4. Hasil Prediksi Klaster EISSN 24069701 (Akuntansi dan Keuangan).V-**Error! Bookmark not defined.**

Tabel V-5. Hasil Prediksi Klaster EISSN 25496050 (International Journal)V-**Error! Bookmark not defined.**

Tabel V-6. Hasil Prediksi Klaster EISSN 25498959 (Jurnal Pendidikan Anak)...V-**Error! Bookmark not defined.**

Tabel V-7. Hasil Prediksi Klaster EISSN 25799258 (Jurnal Pendidikan Matematika) V-**Error! Bookmark not defined.**

Tabel V-8. Hasil Prediksi Klaster EISSN 26141884 (Jurnal Pendidikan)...V-**Error! Bookmark not defined.**

Tabel V-9. Hasil Prediksi Klaster EISSN 26151138 (Jurnal Kesehatan).....V-

Error! Bookmark not defined.

Tabel V-10. Hasil Prediksi Klaster EISSN 26569124 (Jurnal Bisnis dan

Akuntansi).....V-**Error! Bookmark not defined.**

DAFTAR GAMBAR

- Gambar II-1. Arsitektur Klasterisasi Teks (Rozeva & Zerkova, 2017)..... **II-Error! Bookmark not defined.**
- Gambar II-2. Arsitektur BERT (Devlin et al., 2019)**II-Error! Bookmark not defined.**
- Gambar II- 3. Arsitektur metode RUP (Temnenco, 2019)**II-Error! Bookmark not defined.**
- Gambar III-1. RINCIAN KEGIATAN PENELITIAN..... III-7
- Gambar III-2. Kerangka Kerja Penelitian..... III-8
- Gambar IV-1. Persebaran Centroid K-Means.. **IV-Error! Bookmark not defined.**
- Gambar IV-2. Use Case Pengelompokan Artikel Ilmiah Menggunakan IndoBERT dan K-Means..... **IV-Error! Bookmark not defined.**
- Gambar IV-3. Use Case Prediksi Klaster Pengelompokan Artikel Ilmiah Menggunakan IndoBERT dan K-Means . **IV-Error! Bookmark not defined.**
- Gambar IV-4. Antarmuka Perangkat Lunak Memasukkan Data Manual **IV-Error! Bookmark not defined.**
- Gambar IV-5. Antarmuka Perangkat Lunak Memasukkan Data File**IV-Error! Bookmark not defined.**
- Gambar IV-6. Antarmuka Perangkat Lunak Melakukan Klasterisasi**IV-Error! Bookmark not defined.**
- Gambar IV-7. Diagram Activity Melakukan Klasterisasi**IV-Error! Bookmark not defined.**
- Gambar IV-8. Diagram Activity Evaluasi Hasil**IV-Error! Bookmark not defined.**
- Gambar IV-9. Diagram Activity Memasukkan Data**IV-Error! Bookmark not defined.**
- Gambar IV-10. Diagram Activity Melakukan Prediksi Klaster**IV-Error! Bookmark not defined.**
- Gambar IV-11. Diagram Sequence Klasterisasi dan Evaluasi Hasil.....**IV-Error! Bookmark not defined.**

Gambar IV-12. Diagram Sequence Memasukkan DataIV-Error! **Bookmark not defined.**

Gambar IV-13. Diagram Sequence Prediksi KlasterIV-Error! **Bookmark not defined.**

Gambar IV-14. Class Diagram IV-Error! **Bookmark not defined.**

Gambar IV-15. Antarmuka Input Data ManualIV-Error! **Bookmark not defined.**

Gambar IV-16. Antarmuka Hasil Input Data ManualIV-Error! **Bookmark not defined.**

Gambar IV-17. Antarmuka Input Data File..... IV-Error! **Bookmark not defined.**

Gambar IV-18. Antarmuka Hasil Input Data FileIV-Error! **Bookmark not defined.**

Gambar IV-19. Antarmuka Hasil Input Data FileIV-Error! **Bookmark not defined.**

Gambar V-1. Scatter Plot Klasterisasi V-Error! **Bookmark not defined.**

BAB I

PENDAHULUAN

1.1 Pendahuluan

Bab pendahuluan ini akan menguraikan mengenai latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah penelitian, dan sistematika penulisan yang akan menjadi landasan bagi pembahasan di bab berikutnya. Secara keseluruhan, penelitian ini bertujuan untuk menganalisis sebaran topik artikel ilmiah berdasarkan kesamaan konten menggunakan IndoBERT dan K-Means. Penelitian ini diharapkan dapat memberikan wawasan tentang pola distribusi topik artikel ilmiah sehingga mendukung efisiensi pengelolaan literatur ilmiah secara lebih akurat dan terstruktur.

1.2 Latar Belakang

Jumlah artikel ilmiah yang diterbitkan mengalami peningkatan pesat dalam beberapa tahun terakhir, sejalan dengan kemajuan ilmu pengetahuan dan teknologi. Banyaknya artikel ilmiah ini menimbulkan tantangan besar dalam hal pengelolaan dan pengorganisasian artikel-artikel tersebut. Analisis sebaran topik artikel berdasarkan relevansi isi merupakan langkah krusial dalam dunia penelitian untuk meningkatkan efisiensi dalam analisis literatur yang relevan. Akan tetapi, proses pengelompokan atau analisis sebaran secara manual sering kali memakan waktu lama dan berisiko mengalami kesalahan manusia (*human error*). Oleh karena itu, dibutuhkan sistem yang mampu membantu dalam memahami pola distribusi topik artikel ilmiah.

Salah satu teknologi yang banyak digunakan dalam analisis data untuk pengelompokan dan memahami sebaran topik adalah metode K-Means. Algoritma ini berfungsi dengan mengelompokkan data ke dalam beberapa grup berdasarkan kesamaan atribut yang dimiliki. K-Means dapat digunakan untuk mengidentifikasi pola distribusi artikel berdasarkan kesamaan topik atau kata kunci yang muncul dalam teks artikel tersebut (MacQueen, 1967). Algoritma ini dikenal karena kemampuannya dalam memproses data dalam volume besar dengan waktu komputasi yang relatif cepat (Hartigan & Wong, 1979). Selain itu, K-Means juga memiliki implementasi yang sederhana sehingga membuat algoritma ini mudah dipahami dan diaplikasikan, sehingga cocok untuk aplikasi yang membutuhkan analisis pola data yang cepat dan jelas seperti dalam penelitian pasar dan pengelompokan data sosial (Jain, 2010).

Dalam analisis sebaran topik menggunakan K-Means, penggunaan model bahasa adalah hal yang sangat penting. Pada penelitian ini, penulis menggunakan model bahasa IndoBERT yang merupakan salah satu varian dari BERT. BERT telah menjadi salah satu model terkemuka dalam pemrosesan bahasa alami karena kemampuannya untuk menangani tugas-tugas seperti klasifikasi teks, pencocokan kalimat, dan ekstraksi informasi dengan tingkat akurasi yang tinggi. Model ini dilatih menggunakan metode pra-pelatihan dan fine-tuning, yang memungkinkan adaptasi terhadap berbagai jenis tugas bahasa yang berbeda dengan data yang spesifik. BERT dipilih karena mampu menghasilkan embedding teks dengan kualitas tinggi melalui proses encoding yang mendalam pada elemen terkecil dokumen (Devlin et al., 2018). Oleh karena itu, penelitian ini menggunakan model

IndoBERT untuk memahami teks dan K-Means untuk menganalisis pola distribusi topik artikel ilmiah.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah yang dapat diidentifikasi dalam penelitian ini adalah:

1. Bagaimana mengimplementasikan metode IndoBERT dan K-Means untuk menganalisis sebaran topik artikel ilmiah?
2. Bagaimana kualitas hasil analisis sebaran topik artikel ilmiah menggunakan IndoBERT dan K-Means berdasarkan *Silhouette Score*?

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menghasilkan sebuah sistem yang mampu menganalisis sebaran topik artikel ilmiah menggunakan IndoBERT dan K-Means.
2. Menghasilkan kualitas hasil analisis sebaran topik artikel ilmiah menggunakan IndoBERT dan K-Means berdasarkan *Silhouette Score*.

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah:

1. Sistem dapat membantu menganalisis sebaran topik artikel ilmiah berdasarkan konten yang sesuai.

2. Diharapkan penelitian ini dapat menjadi referensi untuk penelitian atau pengembangan selanjutnya.

1.6 Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah:

1. Data yang digunakan berupa data artikel ilmiah berbahasa Indonesia.
2. Data yang digunakan berasal dari 12 jurnal yaitu E-Journal Of Cultural Studies, E-Jurnal Akuntansi, Gadjah Mada International Journal Of Business, International Journal Of Basic And Applied Science, International Journal Of Elementary Education, Jurnal Akuntansi dan Keuangan, Jurnal Bisnis dan Akuntansi, Jurnal Cendekia : Jurnal Pendidikan Matematika, Jurnal Kesehatan Andalas, Jurnal Kesehatan Masyarakat, Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini, Jurnal Pendidikan Teknik Mesin Undiksha.

1.7 Sistematika Penulisan

Sistematika penulisan tugas akhir disusun berdasarkan standar penulisan tugas akhir yang ditetapkan oleh Fakultas Ilmu Komputer Universitas Sriwijaya, yaitu:

BAB I. PENDAHULUAN

Bab ini akan menguraikan latar belakang, rumusan masalah, tujuan dari penelitian, manfaat penelitian, dan batasan masalah. Pokok-pokok yang dibahas dalam bab ini akan menjadi pijakan utama bagi bab selanjutnya.

BAB II. KAJIAN LITERATUR

Bab ini menyajikan landasan teori yang mendukung penelitian. Di dalamnya terdapat tinjauan literatur dan penelitian terdahulu yang relevan dengan penelitian ini, termasuk penjelasan mengenai Model IndoBERT, K-Means, serta penjelasan lain yang terkait.

BAB III. METODOLOGI PENELITIAN

Bab ini akan menguraikan metode dan langkah-langkah yang digunakan dalam penelitian ini mulai dari pengumpulan data, perancangan dari sistem yang dibuat, dan tahapan dalam melakukan penelitian sesuai dengan perancangan.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Bab ini menguraikan proses perancangan perangkat lunak, dimulai dari analisis kebutuhan hingga tahap pengujian untuk mengevaluasi hasil pengembangan perangkat lunak tersebut.

BAB V. HASIL DAN PENELITIAN

Bab ini memaparkan hasil penelitian yang dilaksanakan berdasarkan langkah-langkah dan metode yang telah dirancang. Analisis yang disajikan dalam bab ini akan menjadi dasar dalam merumuskan kesimpulan dari penelitian ini.

BAB VI. KESIMPULAN DAN SARAN

Bab ini memaparkan kesimpulan dan saran dari penelitian yang dilakukan berdasarkan uraian pada bab sebelumnya. Hasil dari penelitian ini diharapkan dapat menjadi bahan referensi atau acuan untuk penelitian selanjutnya dan menghasilkan sistem yang lebih baik lagi.

1.8 Kesimpulan

Bab ini telah menguraikan terkait penelitian yang akan dilakukan mencakup latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, serta batasan masalah dari penulisan yang akan dibuat sebagai dasar pemikiran peneliti.

DAFTAR PUSTAKA

- Chandra, C. M., & Rachmadi, M. (2024). Perbandingan Pengembangan Sistem Dengan Pendekatan Konvensional dan Low-Code Pada Sistem Pendukung Keputusan Penilaian Kinerja Pegawai. *JURNAL PENELITIAN SISTEM INFORMASI (JPSI)*, 2(2), 14–27. <https://doi.org/10.54066/jpsi.v2i2.1729>
- Chang, D.-H., Wang, Y.-H., Hsieh, C.-Y., Chang, C.-W., Chang, K.-C., & Chen, Y.-S. (2021). Incorporating Patient Preferences into a Decision-Making Model of Hand Trauma Reconstruction. *International Journal of Environmental Research and Public Health*, 18(21), 11081. <https://doi.org/10.3390/ijerph182111081>
- Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. <https://doi.org/10.48550/arXiv.1810.04805>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, *abs/1810.04805*. <http://arxiv.org/abs/1810.04805>
- Dewi, D. A. I. C., & Pramita, D. A. K. (2019). *Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali | Matrix: Jurnal Manajemen Teknologi dan Informatika*. <https://ojs.pnb.ac.id/index.php/matrix/article/view/1662>

- Dewi, S., & Pakereng, M. A. I. (2023). IMPLEMENTASI PRINCIPAL COMPONENT ANALYSIS PADA K-MEANS UNTUK KLASTERISASI TINGKAT PENDIDIKAN PENDUDUK KABUPATEN SEMARANG. *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 8(4), Article 4. <https://doi.org/10.29100/jipi.v8i4.4101>
- Dharmawan, S., Mawardi, V. C., & Perdana, N. J. (2023). Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT). *Jurnal Ilmu Komputer Dan Sistem Informasi*, 11(1), Article 1. <https://doi.org/10.24912/jiksi.v11i1.24066>
- Dutta, A., Pal, A., Bhadra, M., Khan, M. A., & Chakraborty, R. (2021). An Improved K-Means Algorithm for Effective Medical Image Segmentation. In J. K. Mandal, S. Mukhopadhyay, A. Unal, & S. K. Sen (Eds.), *Proceedings of International Conference on Innovations in Software Architecture and Computational Systems* (pp. 169–182). Springer. https://doi.org/10.1007/978-981-16-4301-9_13
- Guntara, M., & Lutfi, N. (2023). Optimasi Cacah Klaster pada Klasterisasi dengan Algoritma KMeans Menggunakan Silhouette Coeficient dan Elbow Method. *JuTI “Jurnal Teknologi Informasi,”* 2(1), Article 1. <https://doi.org/10.26798/juti.v2i1.944>
- Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan. *Jurnal Nasional Teknologi dan Sistem Informasi*, 5(1), Article 1. <https://doi.org/10.25077/TEKNOSI.v5i1.2019.17-24>

- Hananto, A. L., Assiroj, P., Priyatna, B., Nurhayati, Fauzi, A., Rahman, A. Y., & Hilabi, S. S. (2021). Analysis Of Drug Data Mining With Clustering Technique Using K-Means Algorithm. *Journal of Physics: Conference Series*, 1908(1), 012024. <https://doi.org/10.1088/1742-6596/1908/1/012024>
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108. <https://doi.org/10.2307/2346830>
- Hediyati, D., & Suartana, I. M. (2021). Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro. *JIEET (Journal of Information Engineering and Educational Technology)*, 5(2), 49–54. <https://doi.org/10.26740/jieet.v5n2.p49-54>
- Hidayat, W. A., & Nastiti, V. R. S. (2024). PERBANDINGAN KINERJA PRE-TRAINED INDOBERT-BASE DAN INDOBERT-LITE PADA KLASIFIKASI SENTIMEN ULASAN TIKTOK TOKOPEDIA SELLER CENTER DENGAN MODEL INDOBERT. *JSiI (Jurnal Sistem Informasi)*, 11(2), Article 2. <https://doi.org/10.30656/jsii.v11i2.9168>
- Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178–210. <https://doi.org/10.1016/j.ins.2022.11.139>
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>

- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP* (No. arXiv:2011.00677). arXiv. <https://doi.org/10.48550/arXiv.2011.00677>
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics: Vol. 5.1* (pp. 281–298). University of California Press. <https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992>
- Muhr, D., Affenzeller, M., & Küng, J. (2023). A Probabilistic Transformation of Distance-Based Outliers. *Machine Learning and Knowledge Extraction*, 5(3), 782–802. <https://doi.org/10.3390/make5030042>
- Paembonan, S., & Abduh, H. (2021). Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat. *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, 6(2), Article 2. https://doi.org/10.51557/pt_jiit.v6i2.659
- Patel, H. (2023). *Customer Segmentation using RFM Analysis and K-Means Clustering to enhance Marketing Strategies; Analysis of Algorithmic Bias in Customer Segmentation Models* [University of Virginia]. <https://doi.org/10.18130/ZY3K-SX64>

- Rianti, R., Andarsyah, R., & Awangga, R. M. (2024). Penerapan PCA dan Algoritma Clustering untuk Analisis Mutu Perguruan Tinggi di LLDIKTI Wilayah IV. *NUANSA INFORMATIKA*, *18*(2), Article 2. <https://doi.org/10.25134/ilkom.v18i2.211>
- Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, *14*(1), e0210236. <https://doi.org/10.1371/journal.pone.0210236>
- Rozeva, A., & Zerkova, S. (2017). Assessing semantic similarity of texts – Methods and algorithms. In *AIP Conference Proceedings* (Vol. 1910). <https://doi.org/10.1063/1.5014006>
- Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, *23*(6), Article 6. <https://doi.org/10.3390/e23060759>
- Situmorang, G. F., & Purba, R. (2024). Deteksi Potensi Depresi dari Unggahan Media Sosial X Menggunakan IndoBERT. *Building of Informatics, Technology and Science (BITS)*, *6*(2), Article 2. <https://doi.org/10.47065/bits.v6i2.5496>
- Suraya, G. R., & Wijayanto, A. W. (2022). Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting: Perbandingan Metode Hierarchical Clustering, K-Means, K-Medoids, dan Fuzzy C-Means dalam Pengelompokan Provinsi di Indonesia Menurut

- Indeks Khusus Penanganan Stunting. *Indonesian Journal of Statistics and Its Applications*, 6(2), Article 2. <https://doi.org/10.29244/ijsa.v6i2p180-201>
- Temnenco, V. (2019). *UML , RUP , and the Zachman Framework: Better together*. <https://www.semanticscholar.org/paper/UML-%2C-RUP-%2C-and-the-Zachman-Framework-%3A-Better-Temnenco/f23cb95b0245f99dda5c951765a23c91087685a9>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Vysala, A., & Gomes, J. (2020). *Evaluating and Validating Cluster Results* (No. arXiv:2007.08034). arXiv. <https://doi.org/10.48550/arXiv.2007.08034>
- Wang, X., Wu, P., Xu, Q., Zeng, Z., & Xie, Y. (2021). Joint image clustering and feature selection with auto-adjointed learning for high-dimensional data. *Knowledge-Based Systems*, 232, 107443. <https://doi.org/10.1016/j.knosys.2021.107443>
- Zubair, Md., Iqbal, MD. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2022). An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. *Annals of Data Science*. <https://doi.org/10.1007/s40745-022-00428-2>