

**ANALISIS KATA KUNCI YANG RELEVAN DENGAN
FAKTOR GAS RUMAH KACA DI INDONESIA
MENGGUNAKAN *TEXT MINING* DAN TEORI LUHN**

LAPORAN TUGAS AKHIR

Oleh:

**Kiagus Muhammad Arsyad
105219002**



**FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER
UNIVERSITAS PERTAMINA
AGUSTUS 2023**



Analisis Kata Kunci yang Relevan dengan Faktor Gas Rumah Kaca di Indonesia Menggunakan Text Mining dan Teori Luhn

Kiaagus Muhammad Arsyad
105219002

*Hak Cipta
Universitas Pertamina*

Dilindungi Undang-Undang

© Copyright of Universitas Pertamina

1. Dilarang mengutip karya tulis ini, kecuali:

- a. menyebutkan sumber sesuai kaidah kecendekaan;
- b. pengutipan hanya untuk keperluan pendidikan, penulisan karya ilmiah atau penelitian;
- c. pengutipan tidak merugikan Universitas Pertamina.

2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.

ANALISIS KATA KUNCI YANG RELEVAN DENGAN FAKTOR GAS RUMAH KACA DI INDONESIA MENGGUNAKAN *TEXT MINING* DAN TEORI LUHN

LAPORAN TUGAS AKHIR

Oleh:

Kiagus Muhammad Arsyad

105219002



**FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER
UNIVERSITAS PERTAMINA
AGUSTUS 2023**

1. Dilarang mengutip karya tulis ini, kecuali:
 - a. menyebutkan sumber sesuai kaidah kecendekian;
 - b. pengutipan hanya untuk keperluan pendidikan, penulisan karya ilmiah atau penelitian;
 - c. pengutipan tidak merugikan Universitas Pertamina.
2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.



LEMBAR PENGESAHAN

Judul Tugas Akhir : Analisis Kata Kunci yang Relevan dengan Faktor Gas Rumah Kaca di Indonesia Menggunakan *Text Mining* dan Teori Luhn
Nama Mahasiswa : Kiagus Muhammad Arsyad
Nomor Induk Mahasiswa : 105219002
Program Studi : Ilmu Komputer
Fakultas : Sains dan Ilmu Komputer
Tanggal Lulus Sidang Tugas Akhir : 3 Agustus 2023

Jakarta, 31 Juli 2023

MENGESAHKAN

Pembimbing I

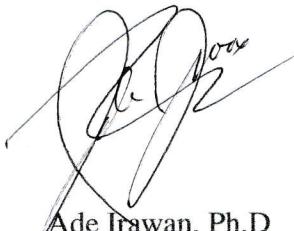
Dr. Tasmi, S.Si., M.Si.
NIP. 116199

Pembimbing II

Drs. Ariana Yunita
NIP. 116015

MENGETAHUI,

Ketua Program Studi



Ade Irawan, Ph.D

NIP. 116130

LEMBAR PERNYATAAN

Dengan ini saya menyatakan bahwa Tugas Akhir berjudul Analisis Kata Kunci yang Relevan dengan Faktor Gas Rumah Kaca di Indonesia Menggunakan *Text Mining* dan Teori Luhn ini adalah benar-benar merupakan hasil karya saya sendiri dan tidak mengandung materi yang ditulis oleh orang lain kecuali telah dikutip sebagai referensi yang sumbernya telah dituliskan secara jelas sesuai dengan kaidah penulisan karya ilmiah.

Apabila dikemudian hari ditemukan adanya kecurangan dalam karya ini, saya bersedia menerima sanksi dari Universitas Pertamina sesuai dengan peraturan yang berlaku.

Demi pengembangan ilmu pengetahuan, saya menyetujui untuk memberikan kepada Universitas Pertamina hak bebas royalti noneksklusif (*non-exclusive royalty-free right*) atas Tugas Akhir ini beserta perangkat yang ada. Dengan hak bebas royalti noneksklusif ini Universitas Pertamina berhak menyimpan, mengalih media/format-kan, mengelola dalam bentuk pangkatan data (*database*), merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya

Jakarta, 31 Juli 2023



Yang membuat pernyataan,

Kiagus Muhammad Arsyad

- b. pengutipan hanya untuk keperluan pendidikan, penulisan karya ilmiah atau penelitian;
- c. pengutipan tidak mengugikan Universitas Pertamina;
2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.



ABSTRAK

Kiagus Muhammad Arsyad. 105219002. Analisis Kata Kunci yang Relevan dengan Faktor Gas Rumah Kaca di Indonesia Menggunakan *Text Mining* dan Teori Luhn.

Penyebab Gas Rumah Kaca (GRK) menjadi isu global yang mendesak untuk ditangani, termasuk di Indonesia. Penelitian ini bertujuan untuk mengidentifikasi faktor penyebab GRK di Indonesia melalui analisis data abstrak dari studi literatur lima tahun terakhir. Pendekatan yang digunakan melibatkan teknik *text mining*, *clustering*, dan seleksi menggunakan teori Luhn. Dengan menggunakan algoritma K-Means dalam teknik *clustering*, data abstrak dikategorikan menjadi beberapa klaster berdasarkan kesamaan konten. Metode seleksi menggunakan teori Luhn juga mampu mengidentifikasi kata-kata kunci yang paling relevan dan mewakili relevansi GRK di Indonesia. Hasil analisis menunjukkan bahwa faktor-faktor pada aspek yang memiliki relevansi GRK di Indonesia meliputi diantaranya deforestasi, biodiesel, CO₂/karbon, dan listrik. Pada penelitian ini juga memberikan pemahaman yang cukup tentang faktor-faktor yang berhubungan pada aspek GRK di Indonesia dan memberikan informasi penting bagi membuat kebijakan, para ahli, para pembaca untuk mengembangkan strategi mitigasi yang efektif.

Kata kunci: faktor, gas rumah kaca, kata kunci, nlp, teori luhn, *text mining*

c. suai kaidah kecendekiaan;

2. Keperluan pendidikan, penulisan karya ilmiah atau penelitian;

c. pengutipan tidak merugikan Universitas Pertamina.

2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.

ABSTRACT

Kiagus Muhammad Arsyad. 105219002. Analysis of Relevant Keywords to Greenhouse Gas Factors in Indonesia Using Text Mining and Luhn's Theory.

The causes of Greenhouse Gas (GHG) emissions have become a pressing global issue to address, including in Indonesia. This study aims to identify the factors contributing to GHG emissions in Indonesia through the analysis of abstract data from literature studies of the past five years. The approach involves techniques such as text mining, clustering, and selection using Luhn's theory. By employing the K-Means algorithm in clustering, abstract data is categorized into several clusters based on content similarity. The Luhn's theory-based selection method also identifies the most relevant keywords that represent the relevance of GHG emissions in Indonesia. The analysis results indicate that factors related to the aspects relevant to GHG emissions in Indonesia include deforestation, biodiesel, CO₂/carbon, and electricity. This study provides a comprehensive understanding of the factors associated with the GHG aspect in Indonesia, offering essential insights for policymakers, experts, and readers to develop effective mitigation strategies.

Keywords: factors, greenhouse gas, keywords, luhn's theory, nlp, text mining



- c. pengutipan tidak merugikan Universitas Pertamina;
- 2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.



KATA PENGANTAR

Dengan memanjatkan segala puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat, taufik, dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi berupa laporan Tugas Akhir (TA) yang berjudul **“Analisis Kata Kunci yang Relevan dengan Faktor Gas Rumah Kaca di Indonesia Menggunakan Text Mining dan Teori Luhn”**, sebagai salah satu persyaratan dalam menyelesaikan masa belajar di Program Studi S1 Ilmu Komputer, Fakultas Sains dan Ilmu Komputer Universitas Pertamina. Penulis menyadari bahwa laporan TA ini tidak akan terselesaikan tanpa adanya dukungan, bimbingan, bantuan, dan nasihat dari berbagai pihak selama penyusunan laporan TA ini. Pada kesempatan ini, penulis ingin menyampaikan terima kasih setulus-tulusnya kepada:

1. Bapak Ade Irawan, Ph.D selaku Ketua Program Studi Ilmu Komputer Universitas Pertamina dan juga sebagai dosen pengaji pertama yang telah memberikan masukan dan saran pada skripsi ini.
2. Bapak Rangga Ganzar Noegraha, Ph.D selaku dosen pengaji kedua yang telah memberikan masukan dan saran pada skripsi ini.
3. Ibu Dr. Tasmi, S.Si, M.Si. dan Ibu Dr. Ariana Yunita selaku dosen pembimbing pertama dan dosen pembimbing kedua pada skripsi ini yang telah memberikan bimbingan, motivasi, dan arahan selama penyusunan laporan tugas akhir dari awal hingga akhir saat ini.
4. Seluruh staf pengajar yang telah memberikan ilmu pengetahuan yang tak ternilai selama penulis menempuh pendidikan di Universitas Pertamina.
5. Seluruh *civitas academica* di lingkungan Universitas Pertamina yang telah menunjang layanan dan dukungan selama proses pembelajaran di Universitas Pertamina.
6. Keluarga, saudara, dan teman-teman penulis yang telah memberikan doa, kasih sayang, motivasi, dan dorongan di setiap langkah perjuangan penulis.

Penulis menyadari bahwa laporan TA ini belum sempurna dan banyak kekurangan, oleh karena itu, segala saran dan kritik yang membangun diharapkan oleh penulis sebagai penyempurnaan penulisan laporan ini di masa yang mendatang. Semoga laporan TA ini dapat memberikan manfaat bagi penulis dan para pembaca.

Jakarta, 31 Juli 2023

Kiagus Muhammad Arsyad

DAFTAR ISI

KATA PENGANTAR	iii
DAFTAR ISI	iv
DAFTAR TABEL	vi
DAFTAR GAMBAR	vii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah	2
1.4 Tujuan penelitian	2
1.5 Manfaat penelitian	2
BAB II TINJAUAN PUSTAKA	4
2.1 Gas Rumah Kaca (GRK)	4
2.2 <i>Text Mining</i>	5
2.2.1 <i>Natural Language Processing</i> (NLP)	5
2.2.2 <i>Term Frequency-Inverse Document Frequency</i> (TF-IDF)	6
2.2.3 <i>Principal Component Analysis</i> (PCA)	7
2.2.4 K-Means	7
2.3 Teknik Kitchenham	7
2.4 <i>Luhn's Theory</i>	8
BAB III METODE PENELITIAN	11
3.1 Tahapan Penelitian	11
3.2 <i>Data Collection</i>	12
3.3 Tahapan <i>Text Mining</i> dan Interpretasinya	14
BAB IV HASIL DAN PEMBAHASAN	17

4.1	<i>Text Preprocessing</i>	17
4.2	Hasil Pengelompokan Kata Kunci berdasarkan teknik <i>clustering</i>	18
4.3	Hasil Pemilihan Kata Kunci berdasarkan Teori Luhn	21
4.4	Hasil Interpretasi Kata Kunci Berdasarkan Teknik <i>Clustering</i> dan Seleksi Teori Luhn	24
BAB V KESIMPULAN DAN SARAN		27
5.1	Kesimpulan	27
5.2	Saran	27
DAFTAR PUSTAKA		29
LAMPIRAN A Lampiran 1. Tren Total Emisi GRK Indonesia		33



DAFTAR TABEL

Tabel 4.1	Total jumlah kata setiap klaster	22
Tabel 4.2	Hasil Kata Kunci yang Terpilih berdasarkan teknik Clustering dan implementasi teori Luhn	24

1. Dilarang mengutip karya tulis ini, kecuali:

- a. menyebutkan sumber sesuai kaidah kecendekiaan;
- b. pengutipan hanya untuk keperluan pendidikan, penulisan karya ilmiah atau penelitian;
- c. pengutipan tidak merugikan Universitas Pertamina.

2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.

DAFTAR GAMBAR

Gambar 2.1	Diagram <i>Word Frequency</i>	9
Gambar 3.1	Diagram Alir Penelitian	11
Gambar 3.2	Tahapan <i>Data Collection</i> hingga hasil interpretasi	12
Gambar 3.3	Tahapan proses pemilihan data studi literatur	13
Gambar 3.4	Alur tahapan saat dilakukan <i>text mining</i> hingga hasil interpretasi	14
Gambar 4.1	Proses tahapan <i>text preprocessing</i> hingga hasil menggunakan data studi literatur	17
Gambar 4.2	Visualisasi hasil dari PCA	19
Gambar 4.3	<i>Elbow Plot</i> untuk menentukan nilai klaster dengan K-Means	20
Gambar 4.4	Metrik Evaluasi nilai K dengan <i>Davies-Bouldin index</i> dan library <i>ElbowVisualizer</i>	20
Gambar 4.5	Klasterisasi Pewarnaan dengan K-Means	21
Gambar 4.6	Representasi visual frekuensi data dengan <i>word cloud</i> pada <i>Cluster 0</i>	22
Gambar 4.7	Representasi visual frekuensi data dengan <i>word cloud</i> pada <i>Cluster 1</i>	22
Gambar 4.8	Representasi visual frekuensi data dengan <i>word cloud</i> pada <i>Cluster 2</i>	23
Gambar 4.9	Representasi visual frekuensi data dengan <i>word cloud</i> pada <i>Cluster 3</i>	23
Gambar 1.1	Total emisi GRK di Indonesia 2010-2019	33





BAB I

PENDAHULUAN

1.1 Latar Belakang

Emisi Gas Rumah Kaca (GRK) di Indonesia diperkirakan meningkat pada periode 2021-2030. Informasi tersebut diperoleh dari artikel DataIndonesia.Id yang ditulis oleh Rizaty [1], pada artikel tersebut juga disebutkan bahwa emisi GRK nasional sudah mencapai 259,1 juta ton CO₂ pada tahun 2021. Proyeksi emisi GRK tahun 2030 diprediksi akan meningkat sebesar 29,13% menjadi 334,6 juta ton CO₂.

Selain itu, Indonesia akan terus meningkat diikuti dengan kemajuan teknologi, hal ini menyebabkan peningkatan kebutuhan energi [2]. Pertumbuhan penduduk ini berdampak pada penggunaan bahan bakar fosil, seperti pembakaran kendaraan bermotor dan kegiatan industri, yang menjadi penyumbang salah satu faktor emisi GRK [3]. Dampak dari peningkatan emisi GRK dan konsumsi energi di dunia juga sangat signifikan terhadap lingkungan, seperti kenaikan suhu global, perubahan iklim ekstrem, serta perubahan pola cuaca [4]. Hal ini perlu diwaspadai mengingat dampak dari emisi GRK yang cukup membahayakan.

Di Indonesia, terdapat program yang dinamakan Indonesia's FOLU Net Sink 2030, yang merupakan sebuah inisiatif dari pemerintah Indonesia untuk mencapai keseimbangan antara emisi dan penyerapan karbon pada tahun 2030. Dalam rangka mengatasi isu perubahan iklim, Indonesia berkomitmen dalam memperkuat sektor kehutanan, pertanian, dan penggunaan lahan lainnya sebagai upaya mencapai "net sink" yaitu penyerapan karbon yang lebih banyak [5].

Inovasi, upaya, dan ide yang dihasilkan dari penelitian yang membahas tentang emisi GRK di Indonesia juga diperlukan saat ini dalam memberikan kontribusi yang lebih efektif dalam mencapai target Indonesia's FOLU Net Sink 2030. Dalam rangka mencari solusi terhadap masalah emisi GRK di Indonesia, pemanfaatan teknologi dan kecerdasan buatan, seperti *Natural Language Processing* (NLP) dan *text mining*, dapat membantu mengatasi tantangan tersebut [6, 7]. Teks-teks dari berbagai sumber, seperti jurnal penelitian, artikel, dan laporan yang membahas mengenai GRK dan terdapat hubungannya di Indonesia dapat dijadikan sebagai data. Data teks tersebut dapat diolah dan dianalisis dengan efisien menggunakan proses *text mining* dan beberapa fungsi dari NLP dalam mengidentifikasi faktor-faktor penyebab GRK yang relevan. Selain itu, terdapat pendekatan menggunakan teori Luhn [8] sebagai *text summarizer* [9] atau dalam pemilihan kata kunci sesuai dengan konsep yang ingin dicari [10], sehingga dapat meningkatkan kualitas dan validitas hasil dari proses tersebut.

Oleh sebab itu, penulis ingin menerapkan pendekatan gabungan berupa kuantitatif dan kualitatif yang dapat menentukan kata-kata kunci yang relevan mengenai GRK di Indonesia berdasarkan data yang tersedia dengan didukung pencarian studi literatur. Diharapkan dari penelitian ini dapat memberikan pengetahuan baru dan ide terkait solusi yang dapat dilakukan untuk mengurangi emisi GRK. Selain itu juga dapat mendukung upaya kepada pembaca, para pakar, peneliti, instansi maupun pemerintah

rintah dalam mengoptimalkan kebijakan dan strategi untuk mengatasi perubahan iklim secara lebih efisien dan efektif khususnya dalam aspek GRK di Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah pada penelitian ini yaitu bagaimana cara menentukan kata-kata kunci yang memiliki relevansi faktor penyebab gas rumah kaca di Indonesia berdasarkan data studi literatur dengan *text mining* berupa teknik *clustering* dan teori Luhn?

1.3 Batasan Masalah

Adapun batasan masalah pada penelitian ini ialah menggunakan data berupa studi literatur yang merupakan artikel atau penelitian lima tahun terakhir sejak tahun 2018 yang didapatkan melalui *search engine* Scopus yang mengulas terkait GRK di negara Indonesia.

1.4 Tujuan penelitian

Tujuan dari penelitian ini yaitu untuk menentukan kata-kata kunci yang memiliki relevansi faktor penyebab gas rumah kaca di Indonesia berdasarkan data studi literatur dengan *text mining* berupa teknik *clustering* dan teori Luhn.

1.5 Manfaat penelitian

Penelitian ini diharapkan dapat memberikan manfaat yang signifikan dalam menentukan faktor penyebab emisi Gas Rumah Kaca (GRK) di Indonesia berdasarkan keterbaruan dari salah satu basis data literatur ilmiah yaitu Scopus. Dengan menerapkan pendekatan gabungan yaitu kuantitatif dan kualitatif menggunakan *text mining* dengan NLP serta K-Means dan teori Luhn, penelitian ini diharapkan juga dapat memberikan kemudahan dan efisiensi dalam mengidentifikasi kata kunci yang relevan mengenai GRK secara valid berdasarkan fakta dan data yang terverifikasi. Selain itu, hasil penelitian ini dapat menjadi dasar untuk mengembangkan kebijakan dan strategi yang lebih baik berdasarkan aspek relevansi dalam menghadapi tantangan perubahan iklim dan menjaga keberlanjutan lingkungan bagi masa depan yang lebih baik terutama dalam kasus GRK di Indonesia.



BAB II

TINJAUAN PUSTAKA

2.1 Gas Rumah Kaca (GRK)

Perubahan iklim telah menjadi isu keamanan internasional yang penting dan nontradisional. Emisi GRK yang berlebihan merupakan salah satu penyebab utama dari meningkatnya permasalahan iklim. Perubahan iklim tidak hanya secara langsung mempengaruhi kesehatan populasi dengan meningkatkan frekuensi dan intensitas gelombang panas, kekeringan, dan curah hujan yang tinggi, tetapi juga secara tidak langsung dengan meningkatkan polusi udara, mempercepat penyebaran vektor penyakit, serta mempengaruhi keamanan pangan dan kesehatan mental. Emisi GRK yang berlebihan menjadi faktor yang memperburuk perubahan iklim dan dampaknya terhadap berbagai aspek kehidupan manusia [11].

GRK merujuk pada gas-gas yang hadir di atmosfer, baik secara alami maupun sebagai hasil aktivitas manusia (antropogenik), yang mampu menyerap dan memancarkan kembali radiasi inframerah [12]. Ketika permukaan bumi menerima radiasi matahari dalam bentuk gelombang pendek, sebagian besar radiasi ini dipancarkan kembali ke atmosfer sebagai radiasi gelombang panjang (infra merah). GRK yang terdapat di lapisan atmosfer yang dekat dengan permukaan bumi menyerap radiasi gelombang panjang ini, menyebabkan peningkatan suhu yang tinggi yang dikenal sebagai efek rumah kaca. Peningkatan suhu ini terjadi akibat perubahan kondisi dan komposisi atmosfer yang mengelilingi planet ini [13].

Dalam era saat ini, masalah lingkungan telah menjadi pembahasan utama baik di negara-negara yang sedang berkembang maupun negara-negara maju karena adanya kerusakan lingkungan. Hal ini juga menimbulkan pertanyaan tentang pemanasan global dan perubahan iklim, yang terutama disebabkan oleh emisi GRK, kadang-kadang terkait dengan penyebab alami seperti pergeseran benua, aktivitas gunung berapi, radiasi matahari, dan arus laut, serta aktivitas manusia langsung maupun tidak langsung yang mempengaruhi komposisi atmosfer global dan variasi lingkungan [4]. Para peneliti telah berargumen bahwa peningkatan aktivitas manusia akibat perkembangan industrialisasi, pertumbuhan populasi global, dan kebutuhan untuk mengatasi perubahan tersebut adalah penyebab utama perubahan iklim. Selain itu, aktivitas manusia seperti deforestasi pertanian dan komersial, pembakaran bahan bakar fosil, serta perubahan penggunaan lahan akibat pertumbuhan populasi juga memberikan kontribusi yang signifikan terhadap peningkatan emisi GRK [14].

Menurut Khairunnisa Musari & Sayah [15], dalam rangka penyelesaian masalah GRK, beberapa hal yang perlu diperhatikan adalah upaya melawan perubahan iklim, prioritas nasional, transformasi kebijakan, menciptakan lingkungan yang mendukung, dan investasi keuangan yang diperlukan, yang semuanya harus menjadi bagian dari agenda nasional. Selain itu dengan perkembangan era informasi saat ini, salah satu upaya dengan pemanfaatan teknologi atau kecerdasan buatan (*Artificial Intelligence/AI*) menjadi hal yang penting karena dapat digunakan untuk pemantauan, analisis, dan pengelolaan data terkait emisi GRK. Sementara itu, dapat membantu dalam mengoptimalkan kebijakan

dan strategi untuk mengurangi emisi GRK secara efisien. Dengan memanfaatkannya secara holistik, diharapkan dapat memberikan penyelesaian masalah GRK dan perubahan iklim dapat tercapai dengan lebih efektif dan efisien.

2.2 *Text Mining*

Studi literatur mengenai *text mining* telah menjadi area penelitian yang semakin populer dalam ilmu data dan pengolahan bahasa alami. *Text mining* merupakan teknik yang digunakan untuk mengekstraksi informasi berharga dari data teks yang tidak terstruktur, seperti artikel jurnal, berita, dan dokumen teks lainnya. Teknik ini mencakup beberapa tahapan, termasuk pengumpulan data teks, pemrosesan teks untuk menghapus karakter-karakter tidak penting, dan analisis data untuk mengidentifikasi pola, topik, atau informasi penting dari teks tersebut [7].

Dalam penerapannya, *text mining* telah digunakan dalam berbagai bidang. Misalnya, dalam analisis sentimen, *text mining* dapat digunakan untuk mengekstraksi sentimen atau perasaan dari teks yang ditulis oleh pengguna dalam media sosial atau ulasan produk [16]. Selain itu, dalam pengelompokan topik, *text mining* dapat membantu mengelompokkan dokumen-dokumen teks berdasarkan tema atau topik yang sama.

Dalam penelitian ini, *text mining* akan digunakan sebagai salah satu teknik dalam *preprocessing dataset* berupa teks untuk mendapatkan kata kunci yang relevan dengan permasalahan GRK di Indonesia. Teknik *text mining* akan membantu dalam ekstraksi fitur teks dari data teks yang tidak terstruktur, yakni abstrak dari jurnal penelitian yang terdapat pada *dataset*. Proses *text mining* akan melibatkan langkah-langkah seperti dengan teknik *stopwords*, *lemmatization*, *punctuation*, *word frequency* dari ekstraksi *unigram* yang akan menjadi kata-kata dasar dari *dataset* tersebut.

2.2.1 *Natural Language Processing (NLP)*

Bahasa Pemrosesan Alami atau yang biasa disebut *Natural Language Processing (NLP)* adalah bidang keilmuan yang berfokus pada pemahaman dan penerapan bahasa manusia dalam bentuk teks atau ucapan oleh komputer. NLP telah mengalami perkembangan pesat dalam beberapa dekade terakhir, berkat kemajuan teknologi komputasi dan kemampuan mesin dalam memproses data secara lebih efektif. Salah satu tantangan utama dalam NLP adalah mengenali dan memahami struktur dan makna dari teks bahasa manusia, karena bahasa manusia seringkali ambigu dan penuh dengan variasi. Beberapa aplikasi praktis dari NLP termasuk mesin penerjemah, asisten virtual, analisis sentimen, pengenalan suara, dan masih banyak lagi. NLP telah menjadi bidang penelitian yang sangat menarik dan relevan dalam era informasi digital yang sedang berkembang pesat [6].

Salah satu pendekatan populer dalam NLP adalah menggunakan teknik pembelajaran mesin (*machine learning*) untuk mengajarkan komputer bagaimana memahami dan memproses bahasa manusia. Teknik-teknik ini melibatkan penggunaan model statistik dan algoritma yang kompleks untuk mengidentifikasi pola, keterkaitan, dan makna dalam teks bahasa manusia. Contoh teknik pembelajaran mesin yang sering digunakan dalam NLP termasuk model berbasis aturan, model berbasis vektor, dan model berbasis jaringan saraf tiruan [17]. Dalam beberapa tahun terakhir, penggunaan jaringan

saraf tiruan, khususnya dalam bentuk transformasi bahasa seperti BERT dan GPT-3, telah mencatat kemajuan signifikan dalam banyak tugas NLP, seperti penerjemahan mesin dan pemahaman bahasa alami [18].

Dalam penelitian ini, teknik pemrosesan bahasa alami (Natural Language Processing/NLP) akan digunakan dalam tahapan *text mining* pada data berupa teks untuk mendapatkan kata kunci yang relevan. Proses pada tahapan tersebut meliputi *preprocessing* dengan menerapkan langkah-langkah seperti menghapus kata umum dan tanda baca, lematisasi, serta mendapatkan kata-kata yang relevan dari teks. Selain itu, teknik ekstraksi fitur seperti metode Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk mengidentifikasi kata-kata yang paling penting dan mewakili konten teks secara numerik [19]. Hasil dari penggunaan NLP dalam pemrosesan data teks ini akan digunakan sebelum menentukan hasil interpretasi yang diharapkan pada penelitian ini.

2.2.2 *Term Frequency-Inverse Document Frequency (TF-IDF)*

TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan salah satu teknik yang sering digunakan dalam pengolahan teks dan *Information Retrieval* (IR). Teknik ini berfungsi untuk memberikan bobot pada setiap kata dalam dokumen berdasarkan frekuensi kemunculan kata tersebut dalam dokumen itu sendiri (TF) dan kebalikannya, yakni berdasarkan keberbedaan jumlah dokumen yang mengandung kata tersebut di dalam seluruh koleksi dokumen (IDF) [20]. Dengan menggunakan TF-IDF, kata-kata yang sering muncul dalam suatu dokumen tetapi jarang muncul di dokumen lain akan memiliki bobot yang tinggi, sehingga dapat dianggap sebagai kata kunci atau kata yang lebih penting dalam konteks dokumen tersebut. Penggunaan TF-IDF telah terbukti efektif dalam berbagai aplikasi seperti klasifikasi teks, analisis sentimen, dan sistem rekomendasi, serta telah menjadi salah satu teknik yang populer dalam mengolah teks untuk mendapatkan informasi yang bermakna [21].

Adapun TF-IDF *Vectorizer* adalah salah satu fungsi yang digunakan dalam pengolahan teks untuk mengimplementasikan metode TF-IDF. TF-IDF *Vectorizer* berfungsi untuk mengubah teks menjadi vektor numerik berdasarkan nilai TF-IDF dari setiap kata dalam teks [21]. Proses ini memungkinkan data teks yang kompleks dan tidak terstruktur diubah menjadi representasi numerik yang dapat diolah lebih lanjut dengan algoritma pembelajaran mesin atau analisis statistik.

Penggunaan TF-IDF *Vectorizer* sangat berguna dalam analisis teks dan NLP, terutama dalam pemrosesan data besar yang mengandung banyak dokumen atau teks. Dengan menggunakan TF-IDF *Vectorizer*, dapat dilakukan suatu pengelompokan teks, analisis sentimen, deteksi topik, dan sistem rekomendasi secara efisien. Selain itu, dengan menerapkan TF-IDF *Vectorizer* dalam penelitian ini mengenai emisi GRK di Indonesia, data teks abstrak akan dikonversi menjadi representasi vektor TF-IDF dengan memperhitungkan bobot kata dalam setiap abstrak. Kata-kata kunci yang relevan dalam data teks yang berkaitan dengan GRK dan menganalisis tingkat signifikansinya dalam dokument tersebut.

2.2.3 Principal Component Analysis (PCA)

PCA (*Principal Component Analysis*) adalah sebuah metode statistik yang digunakan untuk mengurangi dimensi dari data dengan mempertahankan informasi yang paling penting. PCA digunakan untuk mengidentifikasi pola dalam data, dan mengubah data tersebut menjadi bentuk yang lebih mudah dipahami dan diinterpretasikan [22].

PCA lebih sering digunakan dalam analisis data numerik, namun PCA juga dapat digunakan dalam analisis teks. Dalam analisis teks, PCA digunakan untuk mengurangi dimensi dari data teks dengan mempertahankan informasi yang paling penting [23]. PCA dapat digunakan untuk mengidentifikasi pola dalam data teks, seperti pola kata yang sering muncul bersama-sama atau pola topik yang terkait. Dengan mengurangi dimensi dari data teks, PCA dapat mempermudah analisis dan interpretasi data teks.

2.2.4 K-Means

K-Means adalah salah satu algoritma *clustering* yang paling populer digunakan dalam analisis data. Algoritma ini digunakan untuk membagi data menjadi beberapa kelompok atau klaster berdasarkan kesamaan fitur atau karakteristik tertentu [24]. Dalam K-Means, data dikelompokkan berdasarkan jarak antara titik data dan *centroid*, yang merupakan pusat dari setiap klaster. Tujuan dari algoritma ini adalah untuk meminimalkan varians dalam setiap klaster dan memaksimalkan jarak antara klaster.

Salah satu penerapan K-Means yang sering digunakan adalah dalam analisis teks. Dalam analisis teks, K-Means digunakan untuk mengelompokkan dokumen berdasarkan kesamaan topik atau kata kunci tertentu [25]. Misalnya, jika kita memiliki kumpulan dokumen tentang olahraga, K-Means dapat digunakan untuk mengelompokkan dokumen berdasarkan jenis olahraga atau topik tertentu seperti sepak bola, basket, atau tenis. Dalam hal ini, K-Means dapat membantu mengidentifikasi pola dan tren dalam data teks yang besar dan kompleks.

K-Means dapat digunakan untuk memproses hasil dari analisis PCA (*Principal Component Analysis*). PCA adalah teknik statistik yang digunakan untuk mengurangi dimensi data dengan mempertahankan informasi yang paling penting. Setelah PCA diterapkan pada data, K-Means dapat digunakan untuk mengelompokkan data menjadi beberapa klaster berdasarkan kesamaan fitur atau karakteristik tertentu [25]. Hasil dari analisis K-Means dapat membantu mengidentifikasi pola dan tren dalam data yang besar dan kompleks, serta memudahkan interpretasi hasil analisis.

2.3 Teknik Kitchenham

Teknik Kitchenham merujuk pada metode sistematis literature review (SLR) yang dikembangkan oleh Barbara Kitchenham, seorang peneliti terkemuka di bidang rekayasa perangkat lunak [26]. Pedoman Kitchenham untuk melakukan SLR telah menjadi sangat diakui dan sering disebut sebagai "teknik Kitchenham".

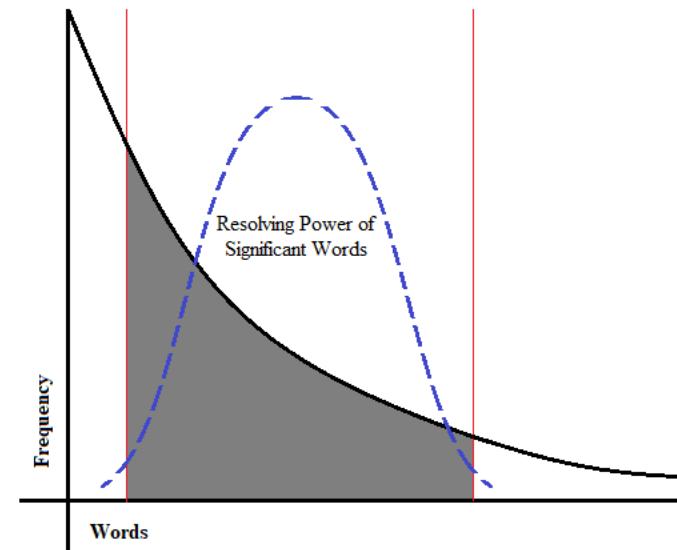
Teknik Kitchenham melibatkan proses dalam mengidentifikasi, mengevaluasi, dan mensintesis semua studi yang relevan pada pertanyaan atau topik penelitian tertentu. Proses ini meliputi pencarian yang komprehensif dan sistematis dari beberapa database dan sumber lainnya, pengembangan kriteria inklusi dan eksklusi, dan teknik analisis data untuk memastikan bahwa hasilnya dapat diandalkan dan tidak bias. Pedoman Kitchenham telah banyak diadopsi dalam penelitian rekayasa perangkat lunak dan telah berkontribusi pada pengembangan rekayasa perangkat lunak berbasis bukti (EBSE) [26, 27]. Teknik Kitchenham telah menjadi metode standar untuk melakukan SLR dalam rekayasa perangkat lunak dan telah membantu meningkatkan kualitas dan keandalan penelitian di bidang tersebut.

Dalam konteks penggunaan teknik Kitchenham, terdapat penggunaan operator logika seperti *AND*, *OR*, dan *NOT* dalam pencarian studi yang relevan. Operator logika tersebut digunakan untuk menggabungkan kata kunci pencarian dan memperluas atau mempersempit cakupan pencarian. Misalnya, jika ingin mencari studi yang relevan dengan topik "*greenhouse gas*" dapat menggunakan operator logika *AND* untuk menggabungkan kata kunci "*greenhouse*" dan "*gas*". Dengan demikian, pencarian akan memperlihatkan studi yang hanya relevan dengan kedua kata kunci tersebut.

Selain itu juga, penggunaan asterisk (*) dalam pencarian studi yang relevan juga dapat mencakup variasi kata yang terkait dengan akar kata tertentu. Misalnya, jika menggunakan kata kunci "*communicat**", pencarian akan mencakup studi yang menggunakan kata "*communicate*", "*communication*", "*communicating*", dan sebagainya. Dengan menggunakan asterisk, akan dapat memperluas cakupan pencarian dan memastikan bahwa semua studi yang relevan ditemukan, terutama jika ada variasi dalam cara menyebutkan kata kunci tertentu.

2.4 *Luhn's Theory*

Teori yang digagas oleh H. P. Luhn, dalam penelitiannya yaitu penciptaan otomatis abstrak literatur menggunakan mesin pemrosesan data [8]. Implementasi Luhn melibatkan analisis teks lengkap dari sebuah artikel dan pemilihan kalimat dan frasa kunci untuk membuat ringkasan singkat dari konten artikel. Mesin pemrosesan data IBM 704 digunakan untuk menghitung ukuran relatif signifikansi untuk kata dan kalimat individu, yang kemudian digunakan untuk memilih informasi paling penting untuk disertakan dalam abstrak. Implementasi ini terbukti efektif dalam menciptakan abstrak yang melayani tujuan abstrak konvensional dan dapat menghemat waktu dan usaha bagi pembaca literatur teknis.



Gambar 2.1. Diagram Word Frequency [8]

Hal yang menjadi sebagai kata kunci dalam penerapan metode otomatis untuk membuat abstrak literatur menggunakan mesin pemrosesan data adalah kata-kata yang paling relevan dan penting dalam artikel yang sedang dianalisis. Kata-kata ini biasanya dipilih berdasarkan topik atau subjek dari artikel tersebut. Dalam *common words* seperti *pronouns*, *prepositions*, dan *articles* dihapus dari daftar kata-kata yang dianalisis karena kata-kata ini tidak memberikan informasi yang signifikan tentang topik atau subjek dari artikel. Sebaliknya, kata-kata yang lebih spesifik dan relevan seperti kata benda dan kata kerja digunakan sebagai kata kunci untuk memilih kalimat-kalimat kunci dalam artikel yang akan digunakan untuk membuat abstrak.

Terdapat penelitian yang dilakukan oleh Rahmah et al [10] dalam menerapkan teori Luhn, dengan Hukum Zipf yang menyatakan bahwa frekuensi kemunculan sebuah kata dalam teks berbanding terbalik dengan peringkatnya dalam daftar frekuensi kata. Dalam konteks pembuatan abstrak otomatis, hukum Zipf dapat digunakan untuk memilih kata-kata kunci yang paling relevan dan penting dalam artikel yang sedang dianalisis. Selain itu, juga membahas tentang skema Bradford [8], yang merupakan metode untuk mengelompokkan jurnal-jurnal ilmiah berdasarkan topik atau subjek yang sama, dengan adanya *lower cut* dan *upper cut* akan dipilih kata-kata yang diseleksi sebagai tema atau konsep yang dicari sebagaimana yang ditunjukkan pada Gambar 2.1 [8]. Metode tersebut digunakan untuk membantu pembaca literatur teknis menemukan jurnal-jurnal yang relevan dengan topik atau subjek yang sedang mereka teliti [10].

1. Dilarang mengutip:
 - a. menyebutkan Universitas Pertamina;
 - b. pengutipan hanya untuk keperluan penelitian;
 - c. pengutipan tidak merugikan Universitas Pertamina.
2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.





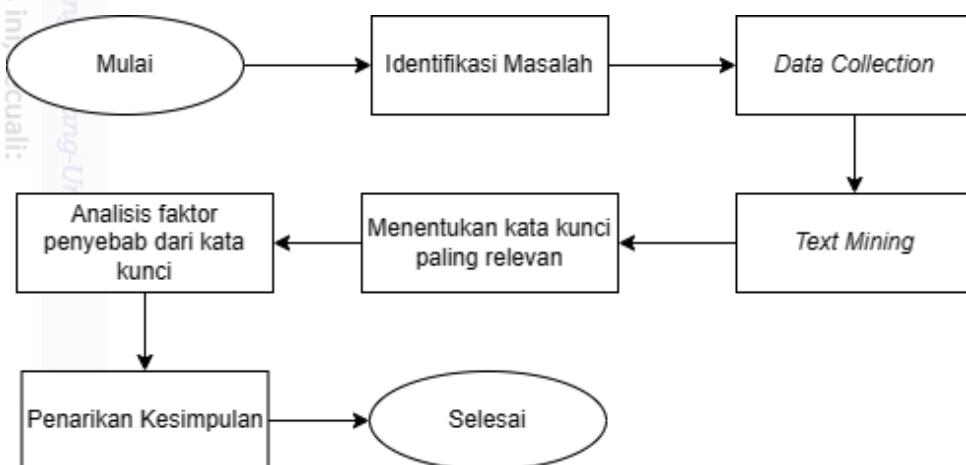
BAB III

METODE PENELITIAN

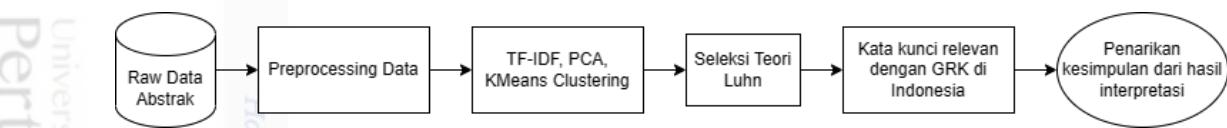
3.1 Tahapan Penelitian

Penelitian ini mengikuti diagram alir yang terdiri dari beberapa tahapan. Tahap pertama adalah mengidentifikasi masalah terkait permasalahan Gas Rumah Kaca (GRK) di Indonesia. Selanjutnya, dilakukan tahap *data collection* dengan mengumpulkan data abstrak dari *database* jurnal penelitian dengan platform Scopus. Setelah itu, data tersebut akan diproses menggunakan teknik *text mining* untuk mendapatkan kata-kata kunci yang relevan.

Selanjutnya, berdasarkan kata-kata dasar yang dihasilkan dari *text mining*, dilakukan pemilihan kata kunci dari kata-kata dasar yang didapatkan dengan mencari relevansi dengan faktor-faktor penyebab GRK, serta hasil analisis menggunakan metode yang dipakai yaitu TF-IDF, PCA, K-Means, dan teori Luhn. Tahap akhir adalah menarik kesimpulan dari hasil penelitian. Dengan demikian, diagram alir penelitian ini menggambarkan langkah-langkah yang sistematis untuk memahami faktor penyebab permasalahan GRK di Indonesia berdasarkan analisis kata kunci.



Pada tahapan ketika melakukan *data collection* hingga ke hasil interpretasi, dengan menerapkan skema yang pernah digunakan oleh Lai [28] dengan skemanya yaitu "*social media to-concepts*", secara rinci akan dijelaskan di alur tahapannya sebagai berikut:



Gambar 3.2. Tahapan *Data Collection* hingga hasil interpretasi

Berdasarkan tahapan yang digambarkan pada Gambar 3.2, hal yang pertama dilakukan yaitu identifikasi masalah terkait emisi GRK di Indonesia yang relevan untuk pencarian data yang dilanjutkan dengan studi literatur pada *dataset* yang mendukung untuk mengidentifikasi faktor-faktor penyebab GRK. Adapun pada pemilihan data akan diambil berdasarkan abstrak yang berhubungan dengan GRK di Indonesia, kemudian dilakukan analisis lebih detil terkait relevansi, topik spesifik, dan keterbaruuan untuk menentukan teknik apa yang akan digunakan dalam pencarian kata-kata kunci dari *dataset* tersebut.

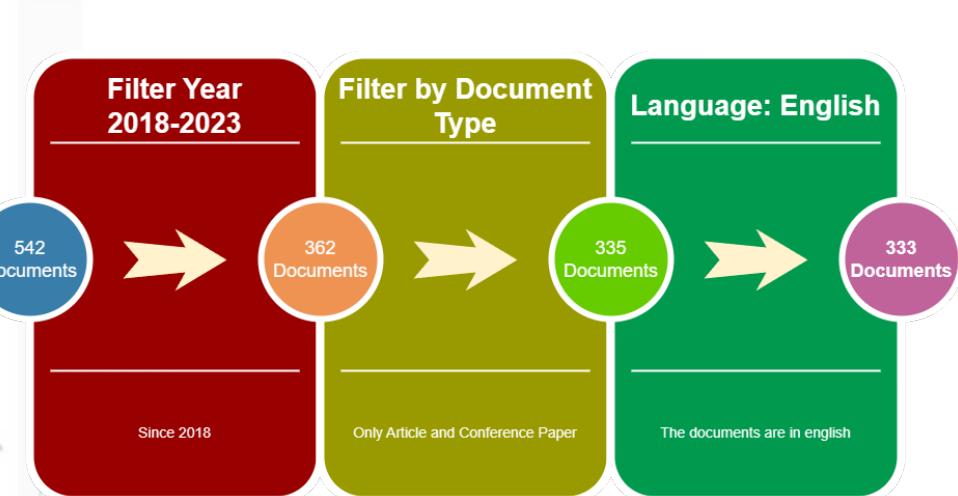
Selanjutnya, data yang telah dikumpulkan akan diproses menggunakan NLP dengan mengimplementasikan *splitting data* secara Unigram (N -grams = 1) untuk mendapatkan kata kunci per huruf dari *dataset*. Kata-kata kunci ini nantinya akan digunakan dalam analisis kualitatif. Dari banyaknya kata kunci tersebut yang memiliki frekuensi paling besar, akan dipilih kata-kata yang paling spesifik berdasarkan teori Luhn. Setelah semua tahapan di atas dilakukan, proses interpretasi akan dilakukan untuk mencari faktor-faktor penyebab GRK berdasarkan hasil yang telah diperoleh. Selanjutnya, akan ditemukan hasil upaya atau solusi dari faktor-faktor penyebab GRK yang telah dipilih berdasarkan kata kunci yang telah didapatkan.

3.2 *Data Collection*

Pada tahapan *data collection*, data yang diambil berdasarkan studi literatur yang telah didapatkan melalui *search engine* Scopus yang mengulas terkait GRK. Data yang diambil tersebut merupakan kumpulan abstrak di artikel atau penelitian yang dipilih berdasarkan pemilihan kata kunci dengan teknik Kitchenham [26]. Dalam mendapatkan data studi literatur, diawali dengan mencari faktor-faktor penyebab GRK terlebih dahulu dengan mendapatkan sumber referensi yang telah pernah membahas faktor penyebab GRK. Adapun cara pencarinya terinspirasi dari penelitian yang dilakukan oleh [29] dengan melakukan teknik Kitchenham [26].

Pada pencarian studi literatur berupa jurnal atau artikel yang akan dipakai sebagai data penelitian ini, dilakukan pencarian di platform Scopus dengan implementasi Kitchenham yakni mendefinisikan objektif penelitian dengan kata kunci agar didapatkan sekumpulan dan abstrak yang relevan secara efektif. Dimulai dari aspek yang berkaitan dengan penyebab, maka kata kunci yang dipakai yaitu "factor, feature, variable, cause, character, impact", kemudian dicari berdasarkan aspek hasil yang ditimbulkan yaitu dengan kata kunci "affect, contribute, product, generate, conduct". Setelah itu, dari segi nama subjek yang dipakai mengenai GRK yakni "GHG, greenhouse gas", dan yang berdasarkan negara atau daerah spesifik yaitu di Indonesia. Maka, rangkaian kata kunci pencarinya yaitu:

(factor OR featur* OR variabl* OR caus* OR character* OR impact) AND (affect* OR contribut* OR produc* OR generat* OR conduc*) AND (ghg OR greenhouse AND gas*) AND (indonesia OR indonesian)*



Gambar 3.3. Tahapan proses pemilihan data studi literatur

Setelah mendefinisikan kata kunci pada pencarian berdasarkan objektif penelitian, terdapat tahapan yang diperlukan dalam pemilihan studi literatur yang cocok sebagaimana yang telah digambarkan pada Gambar 3.3. Pada pencarian pertama, didapatkan sebanyak 542 dokumen yang tersedia. Agar data studi literatur yang dipakai merupakan dokumen yang baru, maka diambil berdasarkan terbitan publikasi 5 tahun saat ini yaitu tahun 2018 hingga saat ini, maka dihasilkan sebanyak 362 dokumen. Setelah itu, untuk mendapatkan jenis studi literatur yang spesifik, maka dipilih jenisnya yaitu *article* dan *conference paper* dan didapatlah sebanyak 335 dokumen. Pada tahap akhir, dari dokumen studi literatur yang didapat perlu dipilih hanya yang berbahasa Inggris saja, untuk pemahaman dan saat pemrosesan data nanti.

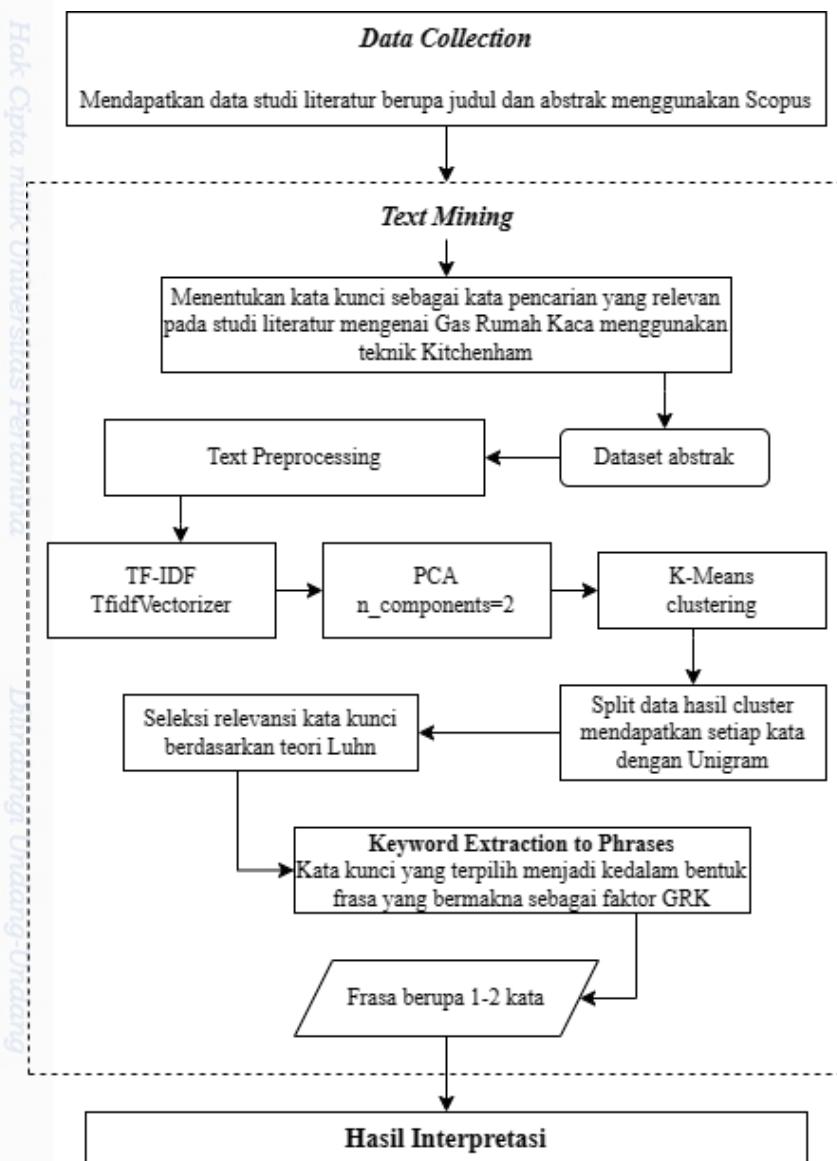
Berikut ini *query search* yang dihasilkan setelah semua pemilihan dokumen studi literatur tersebut dilakukan di platform Scopus:

```
TITLE-ABS-KEY ( factor* OR featur* OR variabl* OR caus* OR character* OR impact AND
affect* OR contribut* OR produc* OR generat* OR conduc* AND ghg OR greenhouse AND gas*
AND indonesia OR indonesian ) AND PUBYEAR > 2017 AND PUBYEAR < 2024 AND ( LIMIT-TO
( LANGUAGE , "English" ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE ,
"cp" ) )
```

Maka, setelah pendefinisian kata kunci dengan teknik Kitchenham, dilanjutkan dengan penyaringan dokumen di platform Scopus, sebanyak 333 dokumen, akan dipakai teks abstrak sebagai *dataset* dalam penelitian ini.

1. Dilarang mengutip atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.
2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya ilmiah dalam penelitian;

3.3 Tahapan *Text Mining* dan Interpretasinya



Gambar 3.4. Alur tahapan saat dilakukan *text mining* hingga hasil interpretasi

Pada tahapan ini, akan lebih banyak implementasi menggunakan NLP, menggunakan data yang berbentuk teks sekumpulan kalimat abstrak yang telah didapatkan, nantinya data tersebut dilakukan *text preprocessing* untuk menyortir teks agar dapat diproses dan siap untuk digunakan dalam analisis lebih lanjut secara baik. Setelah dilakukannya *text preprocessing*, akan dilakukan teknik *clustering* dan dilanjutkan dengan teknik mencari kata kunci yang relevan menggunakan prinsip teori Luhn.

Pada analisis *clustering*, proses tersebut melakukan representasi teks menggunakan TF-IDF, dilanjutkan dengan reduksi dimensi menggunakan PCA, kemudian melakukan K-Means *clustering* untuk mengelompokkan abstrak berdasarkan fitur-fitur yang dihasilkan dari PCA. Hasil akhirnya adalah pengelompokan abstrak ke dalam beberapa klaster berdasarkan kesamaan fitur-fitur TF-IDF yang telah direduksi secara dimensinya menggunakan PCA.

Kemudian pada analisis seleksi kata-kata dasar yang menjadi kata kunci yang memiliki relevansi dengan faktor penyebab GRK, akan diseleksi kata kunci apa yang sesuai berdasarkan cara dari teori Luhn. Nantinya dari data berupa kata kunci yang terpilih atau terseleksi tersebut, serta hasil dari *clustering* sebelumnya, akan dipakai di dalam mencari hasil interpretasi berupa faktor dan subfaktor apa yang sering dibahas sebagai penyebab GRK di Indonesia, dan diakhiri dengan penarikan kesimpulan.

1. Dilarang mengutip karya tulis ini, kecuali:
 - a. menyebutkan sumber sesuai kaidah kecendekiaan;
 - b. pengutipan hanya untuk keperluan pendidikan, penulisan karya ilmiah atau penelitian;
 - c. pengutipan tidak merugikan Universitas Pertamina.
2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.

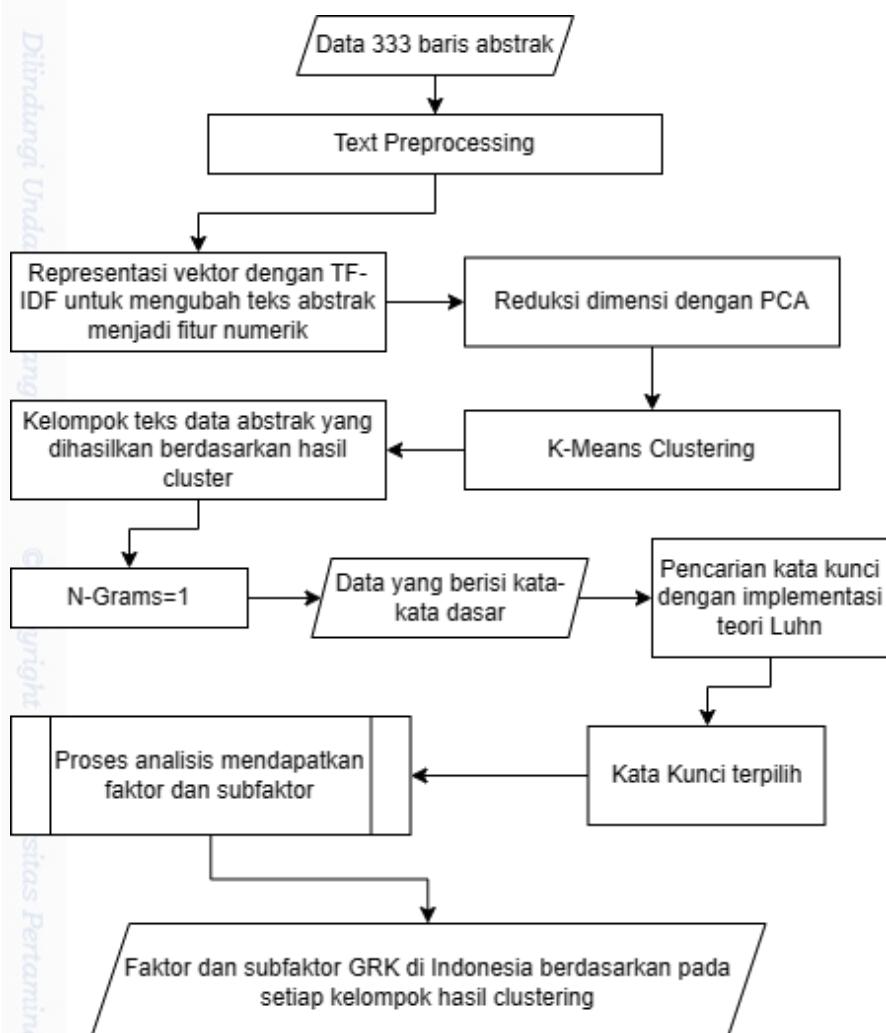




4.1 Text Preprocessing

Data yang telah dikumpulkan berupa 333 dokumen dengan abstrak, pada penelitian ini akan dipakai sehingga berisi 333 baris. Setelah itu, dilakukan suatu *text preprocessing* pada teks tersebut. Pada setelah langkah *text preprocessing* teks tersebut, akan dibagi menjadi dua alur sebagaimana yang telah digambarkan pada Gambar 3.4, yang pertama ialah sebelum dilakukan suatu pengelompokan abstrak serupa menjadi klaster dan yang alur lainnya sebelum dilakukan pemilihan kata dasar untuk mencari kata kunci yang relevan dengan teori Luhn.

Adapun tahapan dalam *text preprocessing* hingga mendapatkan hasil berupa kata kunci yang dicari digambarkan pada gambar di bawah ini:



Gambar 4.1. Proses tahapan *text preprocessing* hingga hasil menggunakan data studi literatur

Di bawah ini, merupakan teknik berupa istilah yang digunakan pada saat melakukan *text preprocessing* sebagai berikut:

1. *Lowercase*: Membuat semua teks pada data tersebut menjadi bukan huruf kapital.
2. *Punctuation*: Menghapus semua unsur tanda baca.
3. *Stopword*: Menghapus unsur kata-kata yang tidak bermakna, seperti '*is, a, it, in, he, is'*.
4. N-Grams=1 (Unigram): Mendapatkan data yang berisi dari kumpulan kata-kata dasar.
5. $\text{len}(\text{word}) > 2$: Mendapatkan kata-kata yang memiliki jumlah huruf lebih dari dua.
6. *Lemmatization*: Memilih satu dari setiap kata dasar yang memiliki bentuk lema yang sama, seperti terdapat kata '*organizing*' dan '*organize*', maka akan dipilih '*organize*'.
7. *No Verb*: Menghapus unsur teks yang mengandung kata kerja.
8. *No Adjective*: Menghapus unsur teks yang mengandung kata sifat.
9. *No Adposition*: Menghapus unsur teks yang mengandung kata kata depan atau kata sambung.
10. *No Auxiliary Verb*: Menghapus unsur teks yang mengandung kata kerja bantu.
11. *No Adverb*: Menghapus unsur teks yang mengandung kata keterangan.
12. *No Geopolitical Entities (GPE)*: Menghapus semua unsur teks yang mengandung unsur nama-nama negara, kota, atau wilayah.
13. *No Numeric*: Menghapus kata-kata yang memiliki unsur angka saja.
14. *No Whitespace*: Menghapus unsur yang hanya berisi kata spasi.

4.2 Hasil Pengelompokan Kata Kunci berdasarkan teknik *clustering*

Pada tahapan awal, data berupa 333 abstrak tersebut dilakukan *preprocessing text*. Kemudian, data yang telah didapatkan tersebut diproses menjadi representasi numerik dari teks abstrak menggunakan metode TF-IDF. TF-IDF digunakan untuk menghitung bobot kata-kata dalam setiap abstrak berdasarkan frekuensi kemunculan kata tersebut di abstrak tertentu dan juga frekuensi kemunculan kata tersebut di semua abstrak. Dengan menggunakan *TfidfVectorizer*, teks abstrak dikonversi menjadi representasi vektor TF-IDF dengan memperhitungkan bobot kata dalam setiap abstrak.

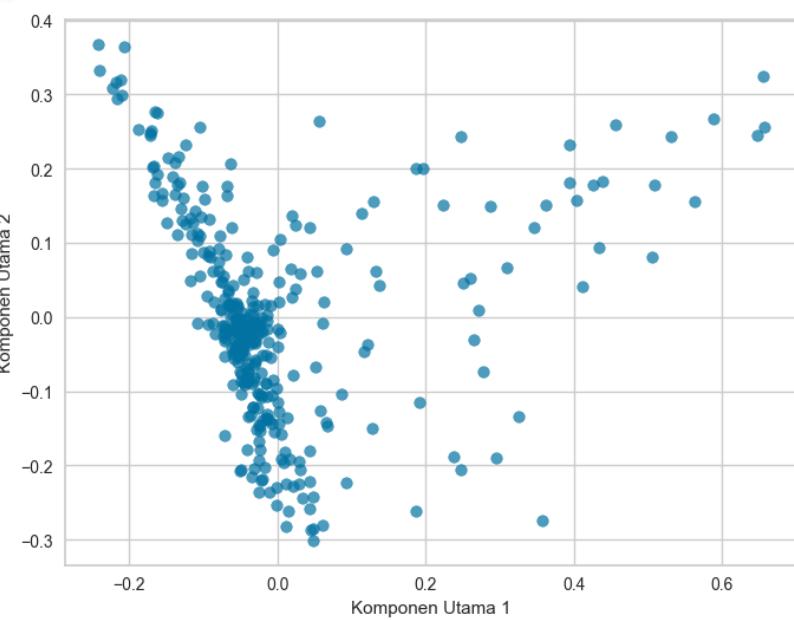
Setelah representasi TF-IDF dibuat, langkah selanjutnya adalah melakukan reduksi dimensi menggunakan PCA. PCA digunakan untuk mengurangi dimensi vektor TF-IDF menjadi dimensi yang lebih rendah, sehingga memungkinkan untuk memvisualisasikan data dalam ruang dua atau tiga dimensi.

Misalnya, terdapat sejumlah fitur yang direpresentasikan sebagai vektor TF-IDF dari teks. Fitur-fitur ini menggambarkan kata-kata dalam abstrak sebagai dimensi dalam ruang vektor. Jika terdapat 4277 kata unik yang dianggap sebagai fitur, maka akan ada 4277 dimensi dalam vektor tersebut.

Kemudian, metode PCA diterapkan untuk mereduksi dimensi-fitur ini. Dalam tahapan sebelumnya, mengakibatkan dimensi ruang vektor dengan 4277 fitur direduksi menjadi hanya 2 komponen utama. Praktiknya, guna mencari dua sumbu utama yang paling mewakili variasi dalam data. Ketika mereduksi dimensi dari 4277 fitur menjadi 2 komponen, sebagian informasi akan hilang. Persentase *variance* yang dijelaskan oleh dua komponen ini akan mengukur besarnya informasi yang tetap dipertahankan.

Dalam perhitungan yang didapatkan, komponen pertama memiliki *variance explained ratio* sekitar 0.0235, yang berarti sekitar 2.35% variasi dalam data oleh komponen ini dan komponen kedua memiliki *variance explained ratio* sekitar 0.0204, yang berarti sekitar 2.04% variasi dalam data dijelaskan oleh komponen ini. Jumlah dari kedua nilai ini adalah sekitar 4.39% sebagai informasi dalam data yang berhasil dipertahankan. Dikarenakan pada penerapannya hanya untuk mereduksi dimensi, informasi yang dipertahankan tersebut tidak akan diterapkan sebagai indikator suatu implementasi pada tahapan mendapatkan kata kunci yang dihasilkan di akhir nanti.

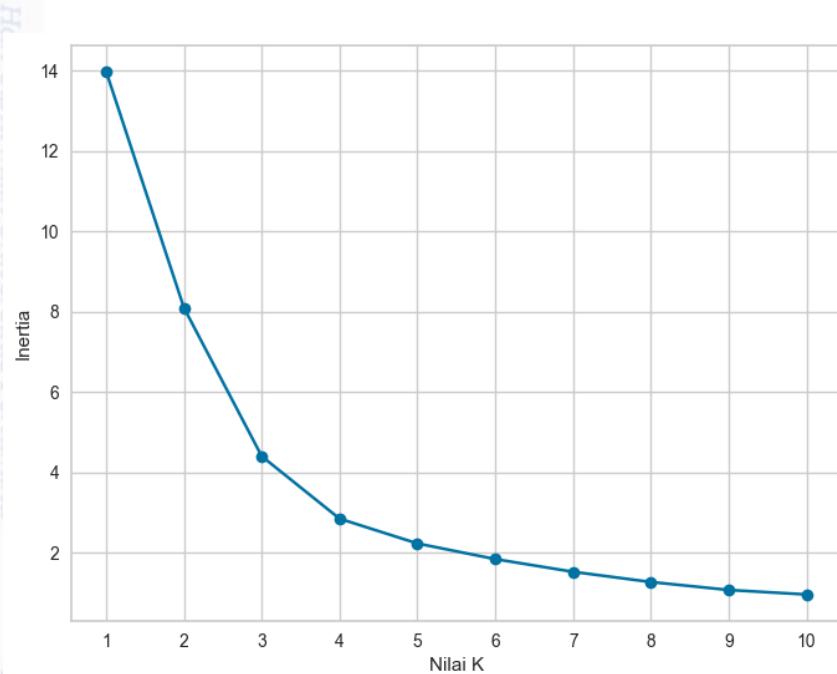
Kembali dalam implementasinya, digunakan $PCA(n_components=2)$ untuk mengambil dua komponen utama dan memvisualisasikan datanya dalam ruang dua dimensi. Komponen utama 1 dan komponen utama 2 pada klasterisasi dengan K-Means merujuk pada hasil reduksi dimensi menggunakan teknik Principal Component Analysis (PCA). Jika nilai komponen utama 1 (nilai x) positif, itu berarti abstrak tersebut cenderung memiliki kontribusi positif dari fitur-fitur yang berkaitan dengan komponen utama 1 tersebut. Sebaliknya, jika nilai komponen utama 1 (nilai x) negatif, itu berarti abstrak tersebut cenderung memiliki kontribusi negatif dari fitur-fitur yang berkaitan dengan komponen utama 1 tersebut.



Gambar 4.2. Visualisasi hasil dari PCA

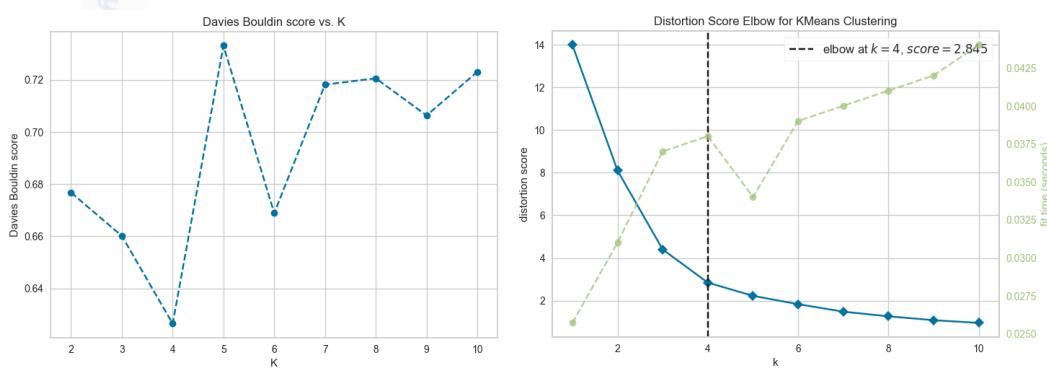
Setelah visualisasi PCA, langkah selanjutnya adalah melakukan K-Means *clustering* untuk mengelompokkan abstrak menjadi beberapa klaster berdasarkan fitur TF-IDF yang telah direduksi menggunakan PCA. Pada implementasi KMeans, dibutuhkan satu parameter yaitu nilai dari K. Untuk

menentukan nilai K atau jumlah *num_clusters* yang optimal, dilakukan pencarian dengan menampilkan *Elbow Plot* menggunakan K-Means, dengan uji coba rentang klaster 1-10.



Gambar 4.3. *Elbow Plot* untuk menentukan nilai klaster dengan K-Means

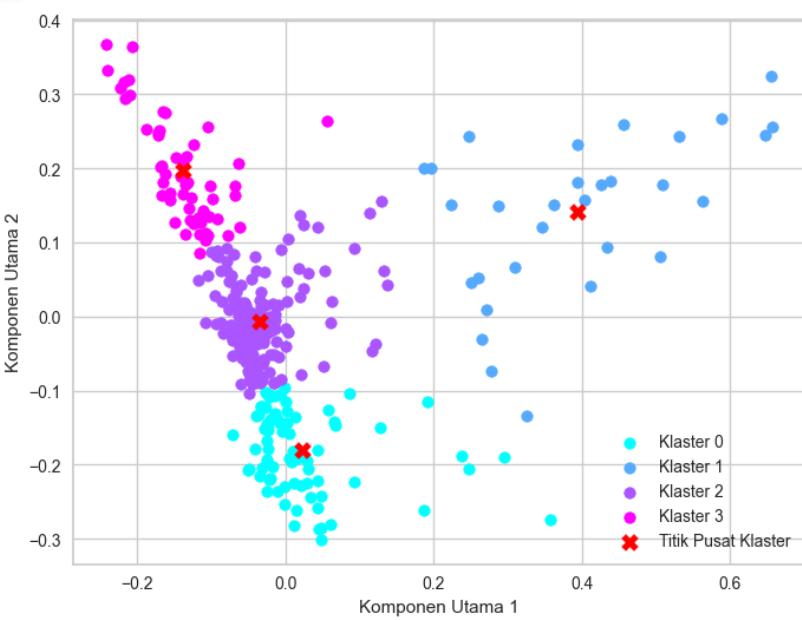
Dalam mempermudah penentuan nilai K yang optimum, dilakukan beberapa metrik evaluasi dengan *Davies-Bouldin index* dan *library ElbowVisualizer*, kedua metrik evaluasi tersebut digunakan untuk mengevaluasi nilai K yang optimal dalam algoritma K-Means. *Davies-Bouldin index* mengukur kualitas pembagian klaster berdasarkan jarak antara klaster yang berdekatan dan dispersi internal setiap klaster [30]. Semakin rendah nilai *Davies-Bouldin index*, semakin baik pembagian klaster dan semakin optimal nilai K yang digunakan. Selain itu, dengan *library ElbowVisualizer* untuk membantu dan memperjelas dari visualisasi *Elbow Plot*, pada saat titik di nilai *inertia* mulai menurun secara lambat atau berhenti menurun drastis, itulah titik yang disarankan sebagai jumlah klaster yang optimal. Hasil dari kedua evaluasi tersebut menunjukkan nilai K=4 yang merupakan nilai K yang optimal, sehingga diterapkan fungsinya yakni *KMeans(n_clusters=4)*.



Gambar 4.4. Metrik Evaluasi nilai K dengan *Davies-Bouldin index* dan *library ElbowVisualizer*

Setelah melakukan K-Means *clustering* dan menentukan jumlah klaster yang optimal berdasarkan *elbow plot* dengan didukung validasi berupa evaluasi metrik, selanjutnya dapat diketahui dalam setiap klaster dengan menggunakan atribut dari objek K-Means untuk menyimpan label klaster pada setiap sampel data abstrak dan mengelompokkannya berdasarkan warna yang berbeda.

Dalam konteks klasterisasi perwarnaan dengan K-Means, komponen utama 1 dan komponen utama 2 adalah dua dimensi baru yang dihasilkan dari PCA. Setiap baris data dalam dataset asli direpresentasikan oleh dua nilai pada komponen utama 1 dan komponen utama 2. Nilai-nilai ini menggambarkan lokasi relatif data dalam ruang dua dimensi yang telah direduksi. Pada saat PCA telah dilakukan, visualisasi data abstrak dalam ruang dua dimensi dengan *scatter plot* menunjukkan bahwa setiap titik pada *scatter plot* merepresentasikan satu abstrak, dan dapat dilihat berupa pola dan struktur data sebelum melakukan *clustering*.



Gambar 4.5. Klasterisasi Pewarnaan dengan K-Means

Pada keempat kluster yang berisi kelompok-kelompok abstrak tersebut, akan dipakai pada tahap selanjutnya, yaitu dilakukan penyeleksian dengan prinsip teori Luhn yang hasilnya nanti akan diterapkan suatu *keyword extraction* untuk menentukan kata-kata kunci yang paling relevan dalam setiap klaster sebagai faktor dan sub-faktor GRK.

4.3 Hasil Pemilihan Kata Kunci berdasarkan Teori Luhn

Pada tahapan memilih kata kunci, berdasarkan teori Luhn, data yang telah didapatkan dari hasil klasterisasi setiap kelompok, dilakukan suatu pemecahan kata-kata dasar pada masing-masing klaster dengan N-Grams = 1 (Unigram).

Untuk mendukung penyelesaian pada tahapan pemilihan kata kunci berdasarkan teori Luhn, dibutuhkan sumber data yang relevan dengan topik penelitian. Dalam hal ini, data yang digunakan berasal dari berbagai sumber terpercaya, seperti jurnal ilmiah, artikel akademik, dan buku referensi. Untuk mendukung penyelesaian pada tahapan pemilihan kata kunci berdasarkan teori Luhn, dibutuhkan sumber data yang relevan dengan topik penelitian. Dalam hal ini, data yang digunakan berasal dari berbagai sumber terpercaya, seperti jurnal ilmiah, artikel akademik, dan buku referensi.

c. pengutipan tidak mengugurkan Universitas Pertamina.

b. menyebutkan sumber sesuai kaidah kecendekiaan;

a. pengutipan hanya untuk keperluan pendidikan, penulisan kary

tanpa adanya izin dari Universitas Pertamina.

Tabel 4.1. Total jumlah kata setiap klaster

Klaster	Total Jumlah Kata
<i>Cluster 0</i>	<i>13198 Kata</i>
<i>Cluster 1</i>	<i>8137 Kata</i>
<i>Cluster 2</i>	<i>2473 Kata</i>
<i>Cluster 3</i>	<i>6476 Kata</i>

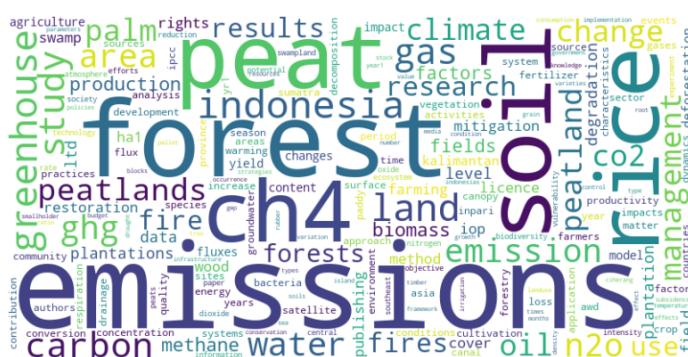
Setelah dilakukan proses pemecahan menjadi kata-kata dasar pada setiap klaster, didapatkan jumlah kata yang didapatkan sebagaimana yang telah ditunjukkan pada Tabel 4.1. Kemudian untuk melihat suatu frekuensi dari masing-masing kata dasar, terdapat visualisasi dengan *Word cloud* untuk menunjukkan representasi visual dari data teks yang menggambarkan frekuensi kata-kata dalam teks tersebut [31]. Dalam *word cloud*, kata-kata yang muncul lebih sering akan ditampilkan dengan ukuran yang lebih besar, sedangkan kata-kata yang muncul lebih jarang akan ditampilkan dengan ukuran yang lebih kecil.

Dilindungi Undang-Undang
mba

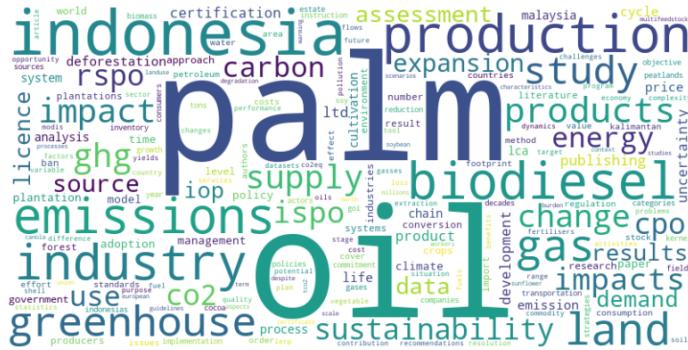
© Copyright of Universitas Pertamina



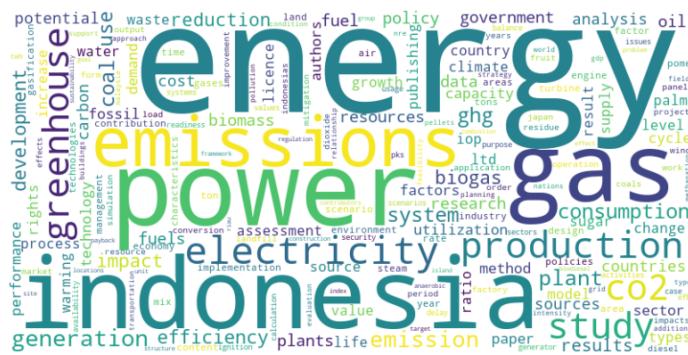
Gambar 4.6. Representasi visual frekuensi data dengan *word cloud* pada *Cluster 0*



Gambar 4.7. Representasi visual frekuensi data dengan *word cloud* pada *Cluster 1*



Gambar 4.8. Representasi visual frekuensi data dengan *word cloud* pada *Cluster 2*



Gambar 4.9. Representasi visual frekuensi data dengan *word cloud* pada *Cluster 3*

Pada keempat representasi visual menggunakan *word cloud* sebelumnya, bisa dilihat bahwa masing-masing setiap klaster memiliki unsur kata dengan jumlah frekuensi yang cukup sedikit berbeda. Terlihat pula kata-kata dasar apa yang memiliki ukuran paling besar dan yang kecil. Namun, pada tahapan seleksi teori Luhn ini bukanlah memilih kata-kata yang memiliki frekuensi terbesar, melainkan semua data dengan jumlah frekuensi kata pada setiap klaster yang akan dipakai pada tahapan teori Luhn. Adapun bagian-bagian kata yang memiliki frekuensi terbesar dan frekuensi terkecil akan dibuang, karena berdasarkan teori Luhn, interval data pada rentang tersebut merupakan istilah-istilah yang kurang baik dalam memilih kata kunci yang relevan secara optimal.

Berikut adalah langkah-langkah dalam pemilihan kata-kata dari hasil *clustering* dan penerapan teori Luhn, serta *labelling* faktor dan subfaktor pada setiap klasternya sebagai berikut:

1. Pemisahan Kalimat Abstrak menjadi Kata-Kata Dasar: Pada setiap klaster, semua kalimat abstrak akan dipecah menjadi kata-kata dasar. Hal tersebut membantu mengidentifikasi kata-kata kunci yang paling mewakili konten dari setiap klaster.
2. Perhitungan Frekuensi Kemunculan Kata: Frekuensi kemunculan setiap kata dalam klaster dihitung. Dengan informasi ini, didapatkan gambaran tentang kata-kata yang paling sering muncul dalam konteks klaster tertentu.
3. Penerapan Teori Luhn: Prinsip teori Luhn digunakan untuk memilih kata-kata yang memiliki frekuensi kemunculan di dalam interval yang ditentukan. Interval ini mengarahkan untuk

2. Dilarang mengutip karya tulis lain, kecuali:

 - a. menyebutkan sumber sesuai kaidah kepercayaan;
 - b. pengutipan hanya untuk keperluan penilaian;
 - c. pengutipan tidak merugikan Universitas Pertamina.

Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.

menghilangkan beberapa kata yang mungkin terlalu umum atau terlalu jarang muncul. Tujuannya adalah mendapatkan kata-kata kunci yang lebih relevan.

4. Penentuan Batas: Berdasarkan teori Luhn, batas interval ditentukan. Batas ini mencakup lower cut (batas bawah) dan upper cut (batas atas). Kata-kata kunci yang masuk dalam interval ini memiliki potensi sebagai representasi yang tepat dari faktor penyebab GRK.
5. Seleksi Kata-Kata Kunci: Kata-kata kunci yang memiliki frekuensi kemunculan di dalam interval batas akan dipilih. Ini menghasilkan kumpulan kata-kata kunci yang akan mewakili konten klaster tertentu.
6. Analisis Kualitatif: Setelah kata-kata kunci dipilih, dilakukan analisis kualitatif lebih lanjut secara manual. Setiap kata kunci akan ditafsirkan dalam konteks faktor penyebab Gas Rumah Kaca (GRK). Hal ini melibatkan pemahaman mendalam tentang makna kata-kata, serta bagaimana kata-kata tersebut terkait dengan aspek GRK.
7. *Labelling* Faktor dan Subfaktor: Setelah analisis kualitatif, kata-kata kunci tersebut dapat digunakan sebagai label untuk mengidentifikasi faktor dan subfaktor GRK dalam setiap klaster. Label ini memberikan representasi komprehensif tentang apa yang mendasari peningkatan GRK.

Setelah melakukan langkah-langkah tersebut pada setiap klaster, maka didapatkanlah faktor dan sub-faktor GRK yang relevan dengan menggunakan kata-kata kunci yang telah dipilih. Proses ini juga merupakan interpretasi akhir mengenai aspek atau faktor penyebab GRK dalam masing-masing konteks klaster.

4.4 Hasil Interpretasi Kata Kunci Berdasarkan Teknik *Clustering* dan Seleksi Teori Luhn

Tabel 4.2. Hasil Kata Kunci yang Terpilih berdasarkan teknik Clustering dan implementasi teori Luhn

Faktor dan 10 Subfaktor Aspek GRK di Indonesia berdasarkan Kata Kunci yang Dihasilkan		
Klaster	Faktor	Subfaktor
Cluster 0	Deforestation	<i>emissions peat, deforestation degradation, mangrove deforestation, mangrove carbon, deforestation mangroves, emissions peatlands, vegetation indices, land fires, emissions peat, forestry land</i>
Cluster 1	Biodiesel	<i>biodiesel production, biomass indonesia, indonesias palm, palm peatlands, biodiesel palm, palm oil, impacts biodiesel, footprint biodiesel, palm expansion, biofuel role</i>
Cluster 2	co2 / Carbon	<i>buildings co2emissions, co2 cement, co2resistant cementing, coal power, waste composition, co2 atmosphere, dependence fuels, contributors carbon, waste business, pollutants countries</i>
Cluster 3	Electricity	<i>electricity consumption, policies energy, buildings energy, boiler efficiency, electricity supply, energy policy, generator efficiency, indonesian electricity, energy sources, capacity biogas</i>

Setelah dilakukan suatu pemilihan kata-kata kunci dari gabungan hasil proses teknik *clustering* dan prinsip teori Luhn tersebut, berdasarkan sumber dari data studi literatur, aspek dan faktor penyebab dari GRK yang ada di Indonesia, maka dari hasil tersebut bisa dijadikan sebagai landasan secara efektif untuk melihat kondisi saat ini dengan memanfaatkan studi literatur 5 tahun terakhir yang membahas terkait subjek atau urgensi perihal GRK di Indonesia.

Contohnya seperti *emissions peat, deforestation degradation, mangrove deforestation* yang merupakan subfaktor dari faktor *deforestation* dan faktor klaster lainnya yang memiliki keterkaitan antara sesama kata-kata kunci yang bisa digabungkan menjadi frasa berupa subfaktor seperti contoh tersebut. Oleh karena itu, dari hasil penelitian ini, dari kata-kata kunci yang didapatkan, dapat diketahui faktor dan subfaktor atau bahkan aspek hal lainnya yang bisa menjadi perhatian bahwa dengan memanfaatkan teknologi saat ini seperti *Artificial Intelligence* dan juga analisis data yang didukung juga berdasarkan teori, dapat membantu dalam analisis tersebut, terutama dalam mengetahui hal-hal yang berhubungan dengan aspek yang dicari.

1. Dilarang mengutip karya tulis ini, kecuali:
 - a. menyebutkan sumber sesuai kaidah kecendekiaan;
 - b. pengutipan hanya untuk keperluan pendidikan, penulisan karya ilmiah atau penelitian;
 - c. pengutipan tidak merugikan Universitas Pertamina.
2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.





BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil analisis data teks menggunakan pengolahan dan teknik dengan *text mining*, *clustering*, dan penyeleksian menggunakan teori Luhn pada data abstrak dari studi literatur lima tahun terakhir, telah didapatkan faktor dan subfaktor kata-kata kunci yang memiliki relevan sebagai faktor penyebab Gas Rumah Kaca (GRK) di Indonesia berupa empat kelompok, yaitu *deforestation*, *biodiesel*, *co2/carbon*, *electricity*. Hasil analisis ini dapat memberikan informasi yang bermanfaat terkait permasalahan GRK saat ini dan dapat menjadi landasan untuk langkah-langkah penanganan dan mitigasi lebih lanjut terhadap dampak GRK di Indonesia terutama pada beberapa aspek yang telah dihasilkan dari keempat kelompok hasil *clustering* sebelumnya.

5.2 Saran

Untuk penelitian selanjutnya, disarankan agar dapat memperluas sumber data studi literatur dengan mencakup lebih banyak publikasi terkini dan diversifikasi sumber literatur, termasuk jurnal ilmiah, laporan pemerintah, dan publikasi akademis lainnya. Penggunaan sumber data yang lebih luas akan memberikan gambaran yang lebih komprehensif terkait faktor penyebab GRK di Indonesia.

Selanjutnya, direkomendasikan untuk melanjutkan penelitian dengan kolaborasi para ahli dalam membantu memvalidasi apa yang telah dihasilkan pada penelitian di bidang terkait. Selain itu, dalam upaya meningkatkan pemahaman yang lebih holistik pada aspek dampak GRK di Indonesia, disarankan untuk dilakukan kajian lebih lanjut dengan mempertimbangkan aspek waktu, geografis, dan perubahan iklim, sehingga dapat memberikan pemahaman yang lebih holistik mengenai dampak GRK di Indonesia yang lebih signifikan dalam pengembangan strategi dan kebijakan mitigasi GRK yang efektif serta berkelanjutan di Indonesia.



DAFTAR PUSTAKA

- [1] M. A. Rizaty, "Emisi Gas Rumah Kaca Indonesia Diproyeksi Terus Naik hingga 2030," *DataIndonesia.Id*, Oct. 2022. [Online]. Available: <https://dataindonesia.id/varia/detail/emisi-gas-rumah-kaca-indonesia-diproyeksi-terus-naik-hingga-2030>
- [2] G. A. Kristanto and W. Koven, "Estimating greenhouse gas emissions from municipal solid waste management in Depok, Indonesia," *City and Environment Interactions*, vol. 4, p. 100027, Dec. 2019. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2590252020300088>
- [3] D. G. K. Ketaren, "PERANAN KAWASAN MANGROVE DALAM PENURUNAN EMISI GAS RUMAH KACA DI INDONESIA," *Jurnal Kelautan dan Perikanan Terapan (JKPT)*, vol. 1, p. 73, Jan. 2023. [Online]. Available: <http://ejournal-balitbang.kkp.go.id/index.php/jkpt/article/view/12050>
- [4] J. Li, M. Irfan, S. Samad, B. Ali, Y. Zhang, D. Badulescu, and A. Badulescu, "The Relationship between Energy Consumption, CO₂ Emissions, Economic Growth, and Health Indicators," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, p. 2325, Jan. 2023. [Online]. Available: <https://www.mdpi.com/1660-4601/20/3/2325>
- [5] M. Fajri, I. Adi Nugroho, and T. Tri Hendro Atmoko Utomo, "Standar Ihk, penunjang target pencapaian folu net sink 2030," *STANDAR: Better Standard Better Living*, vol. 2, no. 3, p. 14–18, Mei 2023. [Online]. Available: <http://majalah.bsilhk.menlhk.go.id/index.php/STANDAR/article/view/124>
- [6] S. P. Tiwari, S. Prasad, and M. Thushara, "Machine learning for translating pseudocode to python: A comprehensive review," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2023, pp. 274–280.
- [7] S. Salloum, M. Al-Emran, A. Monem, and K. Shaalan, *Using Text Mining Techniques for Extracting Information from Research Articles*, 01 2018, pp. 373–397.
- [8] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, Apr. 1958. [Online]. Available: <http://ieeexplore.ieee.org/document/5392672/>
- [9] Computer Education Department, ACLC College of Butuan, Butuan City, Philippines, V. Z. V. Singco, J. C. Trillo, C. C. Abalorio, J. C. M. Bustillo, J. T. Bojocan, and M. C. Elape, "OCR-based Hybrid Image Text Summarizer using Luhn Algorithm with FinetuneTransformer Modelsfor Long Document," *International Journal of Emerging Technology and Advanced Engineering*, vol. 13, no. 2, pp. 47–56, Feb. 2023. [Online]. Available: https://www.ijetae.com/files/Volume13Issue2/IJETAE_0223_07.pdf
- [10] A. Rahmah, H. B. Santoso, and Z. A. Hasibuan, "Critical Review of Technology-Enhanced Learning using Automatic Content Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022. [Online]. Available: <http://thesai.org/Publications/ViewPaper?Volume=13&Issue=1&Code=IJACSA&SerialNo=48>
- [11] H. Wang, J. Luo, M. Zhang, and Y. Ling, "The Impact of Transportation Restructuring on the Intensity of Greenhouse Gas Emissions: Empirical Data from China," *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, p. 12960, Oct. 2022. [Online]. Available: <https://www.mdpi.com/1660-4601/19/19/12960>

- [12] E. Purnamasari, S. Sudarno, and H. Hadiyanto, “INVENTARISASI EMISI GAS RUMAH KACA SEKTOR PERTANIAN DI KABUPATEN BOYOLALI,” *Universitas Muhammadiyah Surakarta*, Apr. 2019.
- [13] R. Pratama, “EFEK RUMAH KACA TERHADAP BUMI,” vol. 14, no. 2, 2019.
- [14] K. O. Yoro and M. O. Daramola, “CO₂ emission sources, greenhouse gases, and the global warming effect,” in *Advances in Carbon Capture*. Elsevier, 2020, pp. 3–28. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/B9780128196571000013>
- [15] K. Musari and F. Sayah, “Green Financing through Green Sukuk in the Fight Against Climate Change: Lessons from Indonesia,” 2021, publisher: Unpublished. [Online]. Available: <http://rgdoi.net/10.13140/RG.2.2.23804.26245>
- [16] A. Aqlan, D. M. Bairam, and R. L. Naik, *A Study of Sentiment Analysis: Concepts, Techniques, and Challenges*, 01 2019, pp. 147–162.
- [17] S. Chung, S. Moon, J. Kim, J. Kim, S. Lim, and S. Chi, “Comparing natural language processing (nlp) applications in construction and computer science using preferred reporting items for systematic reviews (prisma),” *Automation in Construction*, vol. 154, p. 105020, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580523002807>
- [18] S. Moon, S. Chi, and S.-B. Im, “Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (bert),” *Automation in Construction*, vol. 142, p. 104465, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580522003387>
- [19] M. Nasimuzzaman, A. N. Merag, S. Afroz, M. M. Alam, M. H. K. Mehedi, and A. A. Rasel, “Consumer review analysis using nlp and data mining,” in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, March 2023, pp. 0426–0430.
- [20] Y. Setiawan, D. Gunawan, and R. Efendi, “Feature extraction tf-idf to perform cyberbullying text classification: A literature review and future research direction,” in *2022 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, 2022, pp. 283–288.
- [21] K. L. Tan, C. P. Lee, and K. M. Lim, “A survey of sentiment analysis: Approaches, datasets, and future research,” *Applied Sciences*, vol. 13, no. 7, p. 4550, Apr 2023. [Online]. Available: <http://dx.doi.org/10.3390/app13074550>
- [22] T. Kurita, “Principal component analysis (pca),” *Computer Vision: A Reference Guide*, pp. 1–4, 2019.
- [23] J. R. Beattie and F. W. L. Esmonde-White, “Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra,” *Applied Spectroscopy*, vol. 75, no. 4, pp. 361–375, Apr. 2021. [Online]. Available: <http://journals.sagepub.com/doi/10.1177/0003702820987847>
- [24] A. F. Jahwar and A. M. Abdulazeez, “Meta-heuristic algorithms for k-means clustering: A review,” *PalArch’s Journal of Archaeology of Egypt/Egyptology*, vol. 17, no. 7, pp. 12 002–12 020, 2020.
- [25] A. M. Ikorun, A. E. Ezugwu, L. Abualigah, B. Abuaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Information Sciences*, vol. 622, pp. 178–210, Apr. 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0020025522014633>

- [26] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0950584908001390>
- [27] S. Pizard, F. Acerenza, D. Vallespir, and B. Kitchenham, "Assessing attitudes towards evidence-based software engineering in a government agency," *Information and Software Technology*, vol. 154, p. 107101, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950584922002105>
- [28] L. S. L. Lai, "CONTENT ANALYSIS OF SOCIAL MEDIA: A GROUNDED THEORY APPROACH," *Social Media Content Analysis*, vol. 16, no. 2, 2015.
- [29] A. Yunita, H. B. Santoso, and Z. A. Hasibuan, "Finding Contributing Factors of Students' Academic Achievement Using Quantitative and Qualitative Analyses-Based Information Extraction," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 16, pp. 108–125, Aug. 2022. [Online]. Available: <https://online-journals.org/index.php/i-jet/article/view/31945>
- [30] S. Lee, G. Hahn, J. Hecker, S. M. Lutz, K. Mullin, A. D. N. I. (ADNI), W. Hide, L. Bertram, D. L. DeMeo, R. E. Tanzi, C. Lange, and D. Prokopenko, "A comparison between similarity matrices for principal component analysis to assess population stratification in sequenced genetic data sets," *Briefings in Bioinformatics*, vol. 24, no. 1, p. bbac611, 12 2022. [Online]. Available: <https://doi.org/10.1093/bib/bbac611>
- [31] R. M. M. Hicke, M. Goenka, and E. Alexander, "Word Clouds in the Wild," in 2022 IEEE 7th Workshop on Visualization for the Digital Humanities (VIS4DH). Oklahoma City, OK, USA: IEEE, Oct. 2022, pp. 43–48. [Online]. Available: <https://ieeexplore.ieee.org/document/9974812/>
- [32] World Bank, "World development indicators databank," The World Bank, Washington, DC, Tech. Rep., Jul. 2023. [Online]. Available: <https://databank.worldbank.org/source/world-development-indicators>

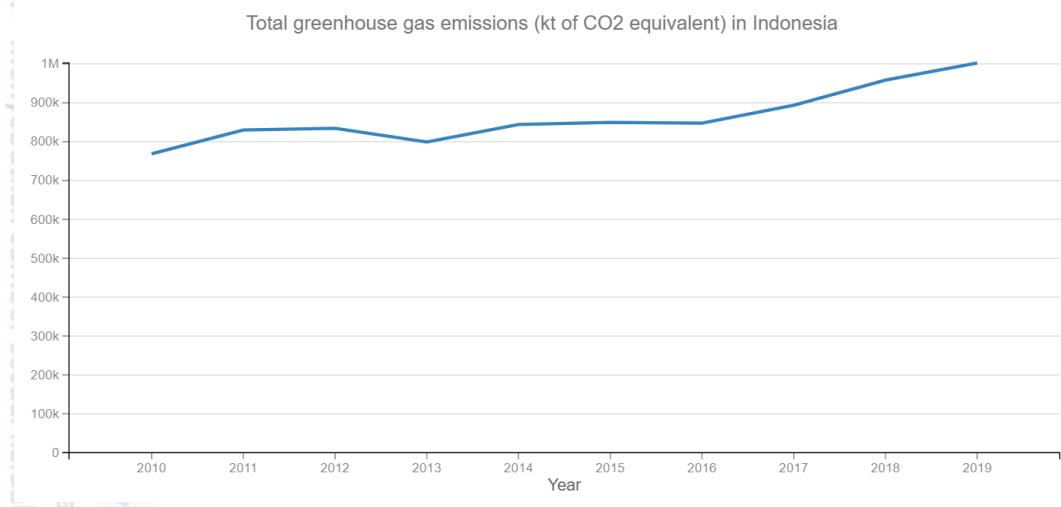




LAMPIRAN A

Lampiran 1. Tren Total Emisi GRK Indonesia

Data yang diambil dari *World Development Indicators* oleh World Bank [32], menunjukkan *total greenhouse gas emissions (kt of CO₂ equivalent)* di Indonesia tahun 2010-2019. Mengindikasikan bahwa tren dari emisi yang ditimbulkan oleh GRK di Indonesia semakin meningkat secara signifikan, pada tahun 2019 dihasilkan sebesar 1.002.369,0 kt of CO₂.



Gambar 1.1. Total emisi GRK di Indonesia 2010-2019

- a. pengutipan hanya untuk keperluan pendidikan, penulisan karya ilmiah atau penelitian;
 - b. pengutipan tidak merugikan Universitas Pertamina;
 - c. pengutipan tidak merugikan Universitas Pertamina.
2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.





Form TA-2 Bimbingan Tugas Akhir
FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER

Nama Mahasiswa : Kiagus Muhammad Arsyad NIM : 105219002
Nama Pembimbing : Dr. Tasmi, S.Si., M.Si. NIP : 116109

No. 1	Hari/Tanggal: Senin / 20 Februari 2023	Paraf Pembimbing
Hal yang menjadi perhatian: Membahas latar permasalahan, ide yang dipakai, dan langkah penulisan selanjutnya. Perlu penekanan terhadap algoritma Machine Learning yang dipakai dalam peningkatan tingkat akurasi atau bisa juga ditampilkan dengan gaya/metode visualisasi yang lebih menarik dan mudah dipahami. Selain itu, perlu ada acuan paper yang bisa dijadikan sebagai referensi/penelitian terdahulu, utamakan cari alasan algoritma yang dipakai dan hasil yang disampaikan pada paper tersebut. Untuk masalah visualisasi dengan web dashboard, itu bisa sebagai poin tambahan untuk menyampaikan hasil visualisasi data agar orang-orang yang membaca bisa juga untuk mencoba dan mudah memahami dengan cara yang menarik serta interaktif.		
No. 2 Hari/Tanggal: Senin / 6 Maret 2023		
Hal yang menjadi perhatian: Pastikan sumber terkait dataset air, mengenai apakah data air tersebut belum diproses atau sudah diproses, kemudian lokasi detil air tersebut. Setelah sudah dapat, baru ke tahap integrasi data dan menambahkan fitur-fitur untuk mendukung data terkait faktor lingkungan dan cuaca. Lanjut jika sudah berhasil, bab1 kelarkan.		
No. 3 Hari/Tanggal: Kamis / 9 Maret 2023		
Hal yang menjadi perhatian: Coba cari terkait data intermittent, apakah bisa diterapkan di GLToy, kemudian, nanti bisa (Tekanan, Temperatur Wellhead, Tekanan Casing di atas permukaan, Suhu di dalam valve, Siklus intermittent gas lift, lamanya injeksi). Ubah topik.		



Form TA-2 Bimbingan Tugas Akhir
FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER

No. 4	Hari/Tanggal: Senin / 15 Mei 2023
Hal yang menjadi perhatian:	
Mengajukan Bab I dengan topik baru dengan judul "Perancangan Dashboard Analisis Preskriptif Personalisasi Profil Pemelajar dengan Pembelajaran Mesin dan Visualisasi Interaktif", mendapatkan saran untuk ditambahkan referensi jurnal internasional.	
No. 5	Hari/Tanggal: Senin / 22 Mei 2023
Hal yang menjadi perhatian:	
<p>"Diberikan saran bagaimana cara membuat penulisan dalam konteks kolaborasi penelitian, selain itu diberi peningkatan dalam permasalahan di latar belakang yang mendasari nya seperti apa serta perbandingan cara yang dipakai.</p> <ul style="list-style-type: none"> - Penyampaian dalam sumber dari pengolahan dari peneliti lainnya - Ikuti referensi mengenai Dashboard ML, terutama langkahnya, metodenya, dan hingga sudah jadinya" 	
No. 6	Hari/Tanggal: Senin / 29 Mei 2023
Hal yang menjadi perhatian:	
Terdapat usulan dalam mengubah tema penelitian mengenai perubahan iklim ke pendekatan Machine Learning atau pemodelan solusi numerik.	
No. 7	Hari/Tanggal: Selasa / 30 Mei 2023
Hal yang menjadi perhatian:	
Ada 2 rujukan paper utama yang bisa dipilih: Paper mengenai mencari faktor/variabel baru yg memengaruhi efek rumah kaca / membuat model lain [Heliyon] ATAU Paper yang meminimalkan emisi CO2 dalam produksi oil and gas serta implementasi suatu metode pada model optimisasi [PS]	



Form TA-2 Bimbingan Tugas Akhir
FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER

No. 8	Hari/Tanggal: Senin / 5 Juni 2023	Paraf Pembimbing:
Hal yang menjadi perhatian:		
	Konsultasi outline Bab 1-4 mengenai rencana penelitian terkait mencari parameter-parameter yang berkaitan sebagai penyebab Gas Rumah Kaca menggunakan algoritma optimisasi. Mendapatkan masukan revisi dan tambahan aspek yang perlu ditulis pada draf Bab 1 dan 2.	
No. 9 Hari/Tanggal: Selasa / 21 Juni 2023		
Hal yang menjadi perhatian:		
	Diskusi dan Ulasan mengenai hasil seminar kemajuan, hal-hal yang perlu direvisi, hal-hal yang perlu dijelaskan dan dikuatkan poin-poin dari skripsi nya. Didapatkan ide inovasi mengarah ke pemodelan menerapkan rumus perhitungan yang terkait dari penelitian peneliti lain.	
No. 10 Hari/Tanggal: Kamis / 13 Juli 2023		
Hal yang menjadi perhatian:		
	Prioritaskan revisi dengan menjawab semua ulasan dari dosen penguji, kemudian kembangkan sesuai dengan inovasi yang diinginkan.	
No. 11 Hari/Tanggal: Senin / 24 Juli 2023		
Hal yang menjadi perhatian:		
	Menelaah kembali progress terakhir draf TA, prioritaskan pendekatan kualitatif lalu kuantitatifnya dengan memastikan fungsi objektif yang valid.	
No. 12 Hari/Tanggal: Rabu / 26 Juli 2023		
Hal yang menjadi perhatian:		
	Memastikan kembali aspek-aspek apa saja yang ditulis pada laporan akhir, bersama pembimbing kedua yaitu bu Ariana, membahas hasil akhir dari laporan yang akan disampaikan dan dikumpulkan pada Tugas Akhir ini.	



Form TA-2 Bimbingan Tugas Akhir
FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER

Nama Mahasiswa : Kiagus Muhammad Arsyad NIM : 105219002
Nama Pembimbing : Dr. Ariana Yunita NIP : 116015

1.	No. 1	Hari/Tanggal: Senin / 27 Februari 2023	Paraf Pembimbing:
c.	Hal yang menjadi perhatian:	Menelaah lagi terkait judulnya pada kata 'Simulasi' karena seperti sebatas simulasi doang gak ada bahasan tuk perbandingan model machine learning nya. Tuk poin-poin besar bab 1 mengenai Sanitasi Air, Teknologi, dan Pembangunan Berkelanjutan tidak masalah. Perkuat hal-hal yang dibuat pada Bab 1 nya, seperti permasalahan kualitas air di beberapa daerah Indonesia, perbandingan akurasi ketepatan model machine learning. Cari juga paper2 yang memodelkan machine learning dan menerapkan web dashboard. Kalau bisa coba cari dataset lain/lagi dan integrasikan data-data tersebut.	
No. 2	Hari/Tanggal: Senin / 6 Maret 2023	Hal yang menjadi perhatian:	Paraf Pembimbing:
	Cari hal ide lain seperti peta Indonesia yang jika diarahkan, memberikan visualisasi seberapa persen terkait permasalahan yang seperti sanitasi air.		
No. 3	Hari/Tanggal: Rabu / 17 Mei 2023	Hal yang menjadi perhatian:	Paraf Pembimbing:
	Menambahkan informasi mengenai personalisasi dan gamifikasi di latar belakang mendapatkan cara berupa Data Mining dari berbagai abstrak dengan kata kunci tertentu di Google Scholar lalu diproses menggunakan Latent Data Allocation.		
No. 4	Hari/Tanggal: Senin / 22 Mei 2023	Hal yang menjadi perhatian:	Paraf Pembimbing:
	Mendapatkan rekomendasi mendapatkan abstrak-abstrak menggunakan Scopus .		



Form TA-2 Bimbingan Tugas Akhir
FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER

No. 5	Hari/Tanggal: Senin / 5 Juni 2023	Paraf Pembimbing:
Hal yang menjadi perhatian:	Setelah adanya topik baru, disarankan agar cari relevansi permasalahan dengan tujuan GRK dengan optimisasi, usahakan tidak berfokus dalam membandingkan antara suatu algoritma karena cukup sulit dan perlu dikaji lebih luas dalam penerapannya. Potensial pertanyaan: Mengapa harus optimisasi, mengapa harus ada istilah optimisasi baru atau lama, mengapa tidak mengimplementasi ML, hasil akhirnya untuk siapa.	
No. 6	Hari/Tanggal: Minggu / 10 Juni 2023	Paraf Pembimbing:
Hal yang menjadi perhatian:	Mendapatkan usulan mengkaji terkait pencarian permasalahan untuk variabel-variabel berdasarkan dataset yang sudah ada. Menambah rumusan dan tujuan tentang hal itu.	
No. 7	Hari/Tanggal: Rabu / 21 Juni 2023	Paraf Pembimbing:
Hal yang menjadi perhatian:	Diskusi dan Ulasan mengenai hasil seminar kemajuan, hal-hal yang perlu direvisi, hal-hal yang perlu dijelaskan dan dikuatkan poin-poin dari skripsi nya. Didapatkan faktor pendukung seperti factors contributing (literature, data), data driven approach, framework pendekatan sektor-sektor, grounded theory.	
No. 8	Hari/Tanggal: Minggu / 23 Juli 2023	Paraf Pembimbing:
Hal yang menjadi perhatian:	Melaporkan hasil progress terkait penggerjaan TA tahap akhir, meliputi metode, revisi, dan ulasan yang telah diajukan sebelumnya.	



Form TA-2 Bimbingan Tugas Akhir
FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER

No. 9	Hari/Tanggal: Selasa / 25 Juli 2023
-------	-------------------------------------

Hal yang menjadi perhatian:

Mempertimbangkan metode terkait tematik analisis dalam pendekatan kualitatif nya.

Paraf Pembimbing:

No. 10	Hari/Tanggal: Rabu / 26 Juli 2023
--------	-----------------------------------

Hal yang menjadi perhatian:

Memastikan kembali aspek-aspek apa saja yang ditulis pada laporan akhir, bersama pembimbing pertama yaitu bu Tasmi, membahas hasil akhir dari laporan yang akan disampaikan dan dikumpulkan pada Tugas Akhir ini.

Paraf Pembimbing:

Universitas Pertamina

2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun

tanpa adanya izin dari Universitas Pertamina.

Dilarang mengutip tulis ini, kecuali:
a. menyajikan kembali sesuai kaidah kecendekiaan;

- b. pengembangan hasil penelitian untuk keperluan pendidikan, penulisan karya ilmiah atau penelitian;

c. penggunaan tidak menguntungkan Universitas Pertamina.

© Copyright of Universitas Pertamina





Universitas
Pertamina

Form TA-4 Pendaftaran Sidang Tugas Akhir
FAKULTAS SAINS DAN ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER

Nama

: Kiagus Muhammad Arsyad

NIM

: 105219002

Judul Tugas Akhir

: Analisis Kata Kunci yang Relevan dengan Gas Rumah Kaca di Indonesia Menggunakan *Text Mining* dan Teori Luhn

Tanggal Seminar Kemajuan TA : 20 Juni 2023

Bersama Form ini disertakan pula:

1. Tiga salinan draft Laporan Tugas Akhir (empat Salinan bila Pembimbing 2 orang)
2. Bukti bahwa Laporan Tugas Akhir telah bebas plagiat
3. Form Bimbingan Tugas Akhir (dengan jumlah bimbingan 8 kali atau lebih)
4. Form bebas pinjam dari laboratorium

Jakarta, 31 Juli 2023

Pendaftar,

Nama : Kiagus Muhammad Arsyad

NIM : 105219002

Menyetujui,

Dosen Pembimbing 1,

Dosen Pembimbing 2,

Nama : Dr. Yasmie, S.Si., M.Si.

NIP : 116109

Nama : Dr. Ariana Yunita

NIP : 116015

Pertamina

1. Dilarang merugikan Universitas Pertamina;
2. Dilarang mempublikasikan atau memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa adanya izin dari Universitas Pertamina.