

**IDENTIFIKASI KATEGORI *SUSTAINABLE DEVELOPMENT GOALS*
(SDGs) DARI PERGURUAN TINGGI BERDASARKAN PUBLIKASI
JURNAL TERINDEKS GARUDA MENGGUNAKAN *BIDIRECTIONAL
ENCODER REPRESENTATIONS FROM TRANSFORMERS* (BERT)**

LAPORAN TUGAS AKHIR

Laporan ini disusun untuk memenuhi salah satu syarat memperoleh Gelar Sarjana Strata 1 (S1) pada Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang



DISUSUN OLEH :

MEISYA REPTIANA

NIM 32602000099

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM SULTAN AGUNG
SEMARANG**

2024

FINAL PROJECT

***IDENTIFICATION OF SUSTAINABLE DEVELOPMENT GOALS (SDGs)
CATEGORIES FROM COLLAGES BASED ON GARUDA INDEXED
JOURNAL PUBLICATIONS USING BIDIRECTIONAL ENCODER
REPRESENTATIONS FROM TRANSFORMERS (BERT)***

*Proposed to complete the requirement to obtain a beachelor's degree (SI)
At Informatics Engineering Department of Industrial Technology Faculty
Sultan Agung Islamic University*



Arranged By :

MEISYA REPTIANA

NIM 32602000099

**MAJORING OF INFORMATICS ENGINEERING
INDUSTRIAL TECHNOLOGY FACULTY
SULTAN AGUNG ISLAMIC UNIVERSITY
SEMARANG**

2024

LEMBAR PENGESAHAN PEMBIMBING

Laporan Tugas Akhir dengan judul "**Identifikasi Kategori Sustainable Development Goals (SDGs) Dari Perguruan Tinggi Berdasarkan Publikasi Jurnal Terindeks GARUDA Menggunakan Bidirectional Encoder Representation From Transformer (BERT)**" ini disusun oleh :

Nama : Meisya Reptiana

NIM : 32602000099

Program Studi : Teknik Informatika

Telah disahkan oleh dosen pembimbing pada :


Hari : **Jumat**


Tanggal : **16 Agustus 2024**

Mengesahkan,

Pembimbing 1


Pembimbing 2


Imam Much Ibnu Subroto, ST., M.Sc., Ph.D
NIK. 210600017


Badie'ah, ST., M.Kom
NIK. 210615044

Mengetahui,

Ketua Program Studi Teknik Informatika
Fakultas Teknologi Industri
Universitas Islam Sultan Agung

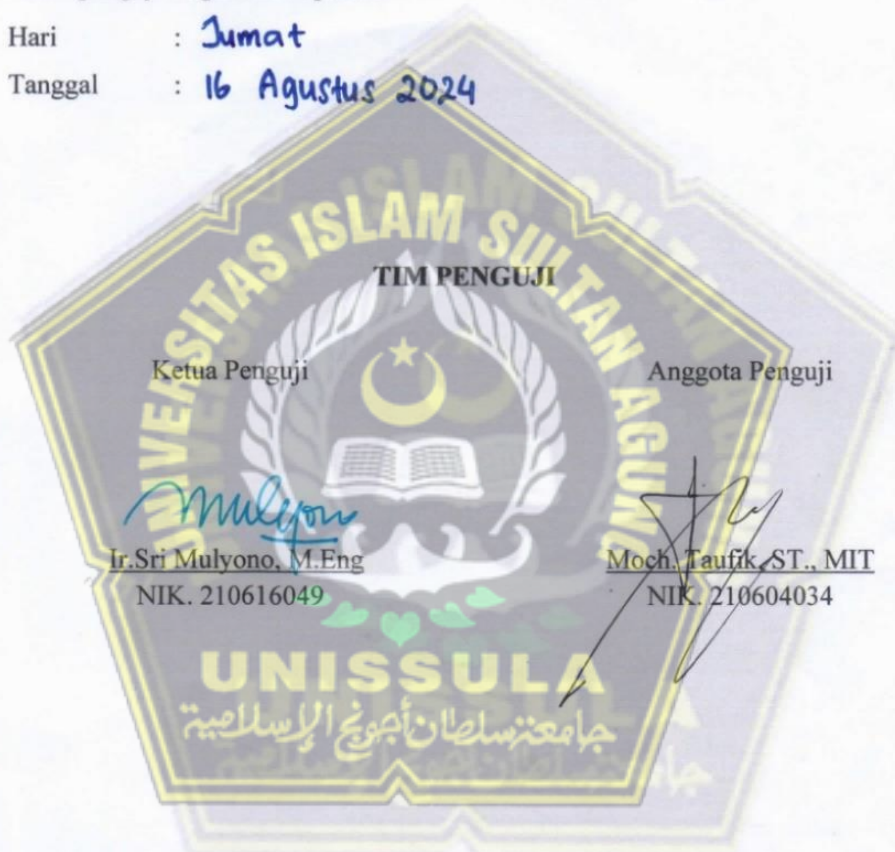

Ir. Sri Mulyono, M.Eng
NIK. 210616049

LEMBAR PENGESAHAN PENGUJI

Laporan tugas akhir dengan judul “*Identifikasi Kategori Sustainable Development Goals (SDGs) Dari Perguruan Tinggi Berdasarkan Publikasi Jurnal Terindeks GARUDA Menggunakan Bidirectional Encoder Representation From Transformer (BERT)*” ini telah dipertahankan di depan dosen penguji Tugas Akhir pada :

Hari : **Jumat**

Tanggal : **16 Agustus 2024**



SURAT PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini :

Nama : Meisya Reptiana

NIM : 32602000099

Judul Tugas Akhir : IDENTIFIKASI KATEGORI *SUSTAINABLE DEVELOPMENT GOALS* (SDGs) DARI PERGURUAN TINGGI BERDASARKAN PUBLIKASI JURNAL TERINDEKS GARUDA MENGGUNAKAN *BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS* (BERT)

Dengan bahwa ini saya menyatakan bahwa judul dan isi Tugas Akhir yang saya buat dalam rangka menyelesaikan Pendidikan Strata Satu (S1) Teknik Informatika tersebut adalah asli dan belum pernah diangkat, ditulis ataupun dipublikasikan oleh siapapun baik keseluruhan maupun sebagian, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka, dan apabila di kemudian hari ternyata terbukti bahwa judul Tugas Akhir tersebut pernah diangkat, ditulis ataupun dipublikasikan, maka saya bersedia dikenakan sanksi akademis. Demikian surat pernyataan ini saya buat dengan sadar dan penuh tanggung jawab.

Semarang, 20 Agustus 2024

Yang Menyatakan,



Meisya Reptiana

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan dibawah ini :

Nama : Meisya Reptiana

NIM : 32602000099

Program Studi : Teknik Informatika

Fakultas : Teknologi industri

Dengan ini menyatakan Karya Ilmiah berupa Tugas akhir dengan Judul : IDENTIFIKASI KATEGORI *SUSTAINABLE DEVELOPMENT GOALS* (SDGs) DARI PERGURUAN TINGGI BERDASARKAN PUBLIKASI JURNAL TERINDEKS GARUDA MENGGUNAKAN *BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS* (BERT)

Menyetujui menjadi hak milik Universitas Islam Sultan Agung serta memberikan Hak bebas Royalti Non-Eksklusif untuk disimpan, dialihmediakan, dikelola dan pangkalan data dan dipublikasikan diinternet dan media lain untuk kepentingan akademis selama tetap menyantumkan nama penulis sebagai pemilik hak cipta. Pernyataan ini saya buat dengan sungguh-sungguh. Apabila dikemudian hari terbukti ada pelanggaran Hak Cipta/Plagiarisme dalam karya ilmiah ini, maka segala bentuk tuntutan hukum yang timbul akan saya tanggung secara pribadi tanpa melibatkan Universitas Islam Sultan agung.

Semarang, 20 Agustus 2024

Yang menyatakan,



Meisya Reptiana

KATA PENGANTAR

Segala puji dan syukur tercurahkan kepada Allah SWT., yang telah memberikan rahmat dan karunianya sehingga penulis dapat menyelesaikan laporan Tugas Akhir dengan judul “Sistem Identifikasi Kategori *Sustainable Development Goals* (SDGs) Dari Perguruan Tinggi Berdasarkan Publikasi Jurnal Terindeks GARUDA” untuk memenuhi salah satu syarat menyelesaikan studi dan memperoleh gelar sarjana (S-1) DI Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung.

Selaku penulis, saya ingin mengucapkan terima kasih kepada berbagai pihak yang telah membantu dalam penyusunan dan pembuatan tugas akhir ini, yaitu :

1. Rektor UNISSULA Bapak Prof. Dr. H. Gunarto, S.H., M.Hum;
2. Dekan Fakultas Teknologi Industri UNISSULA Ibu Dr. Ir. Novi Marlyana, ST., MT., IPU., ASEAN.Eng;
3. Kaprodi Teknik Informatika UNISSULA Bapak Ir. Sri Mulyono, ST., M.Eng;
4. Dosen pembimbing I, Bapak Imam Much Ibnu Subroto, S.T., M.Sc., Ph.D dan Dosen Pembimbing II, Ibu Badie'ah, S.T., M.Kom, yang telah meluangkan waktunya dan memberikan ilmu serta saran kepada penulis;
5. Orang tua dan keluarga penulis yang telah memberikan doa dan dukungan penuh dalam menyelesaikan tugas akhir ini;
6. Teman-teman penulis yang selalu menemani dan membantu setiap proses pengerjaan tugas akhir ini.

Dengan kerendahan hati, penulis menyadari bahwa laporan ini masih memiliki banyak kekurangan dalam hal kualitas, kuantitas, dan ilmu pengetahuan. Oleh sebab itu, penulis mengharapkan kritik dan saran yang membangun untuk membantu memperbaiki laporan ini.

Semarang, 31 Juli 2024

Meisya Reptiana

DAFTAR ISI

HALAMAN JUDUL	i
LEMBAR PENGESAHAN PEMBIMBING	iii
LEMBAR PENGESAHAN PENGUJI.....	iv
SURAT PERNYATAAN KEASLIAN TUGAS AKHIR.....	v
PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	vi
KATA PENGANTAR.....	vii
DAFTAR ISI.....	viii
DAFTAR TABEL	x
DAFTAR GAMBAR	xi
ABSTRAK	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Perumusan Masalah.....	3
1.3 Pembatasan Masalah	3
1.4 Tujuan.....	3
1.5 Manfaat.....	4
1.6 Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA DAN DASAR TEORI.....	5
2.1 Tinjauan Pustaka.....	5
2.2 Dasar Teori	7
2.2.1 <i>Sustainable Development Goals (SDGs)</i>	7
2.2.2 <i>Natural Language Processing (NLP)</i>	8
2.2.3 <i>Multi-Class Classification</i>	8
2.2.4 <i>Data Pre-processing</i>	9
2.2.5 <i>Synthetic Minority Oversampling Technique (SMOTE)</i>	10
2.2.6 <i>Transformer</i>	11
2.2.7 <i>Bidirectional Encoder Representations from Transformer (BERT)</i> 15	
2.2.8 <i>Evaluasi Performa Model</i>	18
BAB III METODE PENELITIAN	19
3.1 Tahapan Penelitian.....	19

3.1.1	Studi Literatur	19
3.1.2	Pengumpulan Data	19
3.1.3	Perancangan Model.....	20
3.1.4	Implementasi Streamlit	25
BAB IV HASIL DAN ANALISIS PENELITIAN.....		28
4.1	Hasil Perancangan Model.....	28
4.1.1	<i>Data Preprocessing</i>	28
4.1.2	<i>Balancing Data Training</i>	33
4.1.3	Hasil Pelatihan Model.....	34
4.2	Hasil Evaluasi.....	34
4.3	Hasil Implementasi Steamlit.....	36
4.3.1	Tampilan Awal Sistem Identifikasi.....	36
4.3.2	Halaman Hasil Sistem Identifikasi.....	37
4.3.3	Halaman Hasil Identifikasi.....	39
4.3.4	Hasil Tampilan <i>Error</i>	40
4.4	Hasil Pengujian Sistem.....	40
BAB V KESIMPULAN DAN SARAN		44
5.1	Kesimpulan.....	44
5.2	Saran.....	44
DAFTAR PUSTAKA		

DAFTAR TABEL

Tabel 3. 1 Distribusi antar kelas data training.....	22
Tabel 4. 1 Distribusi data untuk setiap kelas.....	29
Tabel 4. 2 Perbandingan sebelum dan sesudah data cleaning.....	31
Tabel 4. 3 Proses Tokenisasi	32
Tabel 4. 4 Konfigurasi hyperparameter dengan grid search	34
Tabel 4. 5 Hasil evaluasi model	34
Tabel 4. 6 Hasil evaluasi setiap kelas data testing	36
Tabel 4. 7 Hasil Pengujian Sistem Identifikasi	41



DAFTAR GAMBAR

Gambar 2. 1	Arsitektur Transformer(Al-Faruq, 2021).....	11
Gambar 2. 2	Scaled Dot-Product (kiri) dan Multi-Head Attention (kanan)(Al-Faruq, 2021).....	12
Gambar 2. 3	Arsitektur BERT dengan 12 blok encoder.....	15
Gambar 2. 4	Pre-training dan fine tuning BERT(Devlin dkk., 2019).....	16
Gambar 2. 5	Arsitektur dan Komponen DistilBERT(Putri dkk., 2023).....	17
Gambar 3. 1	Flowchart metode penelitian.....	19
Gambar 3. 2	Flowchart alur pemodelan sistem.....	20
Gambar 3. 3	Flowchart data pre-processing.....	21
Gambar 3. 4	Flowchart inferensi model.....	24
Gambar 3. 5	Rancangan halaman awal sistem identifikasi.....	25
Gambar 3. 6	Rancangan halaman hasil sistem identifikasi.....	26
Gambar 4. 1	Informasi dataset.....	28
Gambar 4. 2	Sampel dataset.....	29
Gambar 4. 3	Hasil data cleaning.....	30
Gambar 4. 4	Jumlah data hasil balancing.....	33
Gambar 4. 5	Hasil Confision Matrix.....	35
Gambar 4. 6	Halaman awal streamlit.....	36
Gambar 4. 7	Hasil identifikasi SDGs 4.....	37
Gambar 4. 8	Hasil identifikasi SDGs 3.....	38
Gambar 4. 9	Hasil identifikasi kategori others.....	38
Gambar 4. 10	Hasil tampilan error.....	40

ABSTRAK

Sustainable Development Goals (SDGs) merupakan komitmen global yang disepakati seluruh Negara Anggota Perserikatan Bangsa-Bangsa (PBB) sebagai agenda tahun 2030, untuk mewujudkan kehidupan dunia yang lebih baik dan berkelanjutan. Hal tersebut, menjadikan Indonesia yg merupakan salah satu negara anggota PBB harus ikut berkomitmen dalam mewujudkan tercapainya 17 tujuan yang ada dalam SDGs. Perguruan Tinggi memiliki peranan penting dalam tercapainya tujuan SDGs, yang dapat diukur melalui penelitian dan publikasi ilmiah. GARUDA merupakan portal *repository* yang memberikan akses terhadap berbagai sumber pengetahuan ilmiah, termasuk di dalamnya jurnal-jurnal dari perguruan tinggi di Indonesia. Untuk mengetahui sejauh mana peranan perguruan tinggi, dapat dilakukan identifikasi kategori SDGs terhadap publikasi jurnal yang dilakukan dan telah terindeks GARUDA. Untuk itu, penelitian ini bertujuan mengembangkan sistem identifikasi kategori SDGs menggunakan model DistilBERT yang merupakan versi distilasi dari BERT. Dari proses pelatihan model yang dilakukan, memperoleh hasil *accuracy* sebesar 86,73% terhadap data *training*. Selanjutnya dari pengukuran metrik evaluasi terhadap data *testing*, diperoleh hasil untuk *accuracy* sebesar 82,56%, *precision* 82,94%, *recall* 82,56%, dan *f1-score* 82,57%. Hasil ini menunjukkan model memiliki performa yang cukup baik dalam mengidentifikasi kategori SDGs berdasarkan data abstrak publikasi jurnal yang terindeks GARUDA.

Kata Kunci : sistem identifikasi, SDGs, BERT, DistilBERT, GARUDA

ABSTRACT

The Sustainable Development Goals (SDGs) are the global commitments agreed by all United Nations member states as the 2030 agenda, to a better and more sustainable world. This, making Indonesia one of the United Nations member states, must engage in achieving the 17 goals contained in the SDGs. Colleges have an important role to play in achieving the goals of the SDGs, which can be measured through research and scientific publication. GARUDA is a repository portal that provides access to various sources of scientific knowledge, including journals from colleges in Indonesia. For this purpose, the study aims to develop a system of identification of categories of SDGs using the DistilBERT model, which is a distillation version of BERT. From the model training process carried out, obtained an accuracy result of 86.73% against data training. Further from the metric measurement of evaluation against data testing, received results for accuracy of 82.56%, precision 82.94%, recall 82.56%, and f1-score 82.57%. These results showed that the model has a fairly good performance in identifying SDG categories based on the abstract data of public journals indexed by GARUDA.

Keywords: identification systems, SDGs, BERT, DistilBERT, GARUDA

BAB I PENDAHULUAN

1.1 Latar Belakang

Sustainable Development Goals (SDGs) merupakan komitmen global yang disepakati seluruh Negara Anggota Perserikatan Bangsa-Bangsa (PBB) sebagai agenda tahun 2030, untuk mewujudkan kehidupan dunia yang lebih baik dan berkelanjutan. *Sustainable Development Goals* (SDGs) memiliki 17 tujuan berkaitan dengan hampir semua aspek kehidupan, yang dapat tercapai melalui adanya kontribusi dari berbagai pihak. Perguruan Tinggi menjadi salah satu pihak yang harus ikut berkontribusi sebagai institusi yang bergerak di bidang pendidikan.

Peran perguruan tinggi sangat penting dalam mendukung SDGs karena melalui berlangsungnya proses pendidikan yang berkualitas dapat membawa manfaat untuk pembangunan berkelanjutan secara signifikan bagi individu, komunitas, hingga negara. Kontribusi perguruan tinggi diperlukan secara luas dalam menciptakan dan menyebarkan pengetahuan, sehingga perguruan tinggi dapat mendorong adanya inovasi global, nasional dan lokal, pembangunan ekonomi, hingga kesejahteraan masyarakat, sebagai cakupan tercapainya agenda SDGs (Mellyana, 2021). Fungsi utama dari perguruan tinggi adalah memberikan pengetahuan dan solusi untuk mendukung implementasi SDGs. Perguruan tinggi memiliki peran besar dalam mendorong kemajuan teknologi dan masyarakat, melalui penelitian, penemuan, dan penciptaan pengetahuan. Maka dari itu, perlu untuk mengetahui sejauh mana peranan perguruan tinggi melalui upaya berkelanjutan untuk mencapai seluruh agenda SDGs. Bukti komitmen perguruan tinggi dalam mendukung terwujudnya SDGs, dapat dilihat melalui penelitian dan penemuan dari berbagai disiplin ilmu yang direpresentasikan dalam bentuk tulisan dan kemudian dipublikasikan dalam jurnal.

Science and Technology Index (SINTA) merupakan *database* publikasi jurnal yang yang terbesar di Indonesia. SINTA memiliki beberapa fitur seperti *Citation*, *Research Output*, dan *Networking* yang dapat memudahkan pengelolaan jurnal sebagai pengindeks global secara internasional (Suryaningsum, 2020). Berdasarkan

pada *website* resmi SINTA, terdapat 281.649 *authors*, 5.599 *affiliations*, dan 2.199 *departments* yang telah melakukan publikasi. Sementara itu, GARUDA merupakan portal *repository* yang memberikan akses terhadap berbagai sumber pengetahuan ilmiah, termasuk di dalamnya beberapa jurnal terindeks SINTA. SINTA hanya memuat data profil bibliografi dari penelitian nasional, sedangkan GARUDA memuat *detail* seperti judul, abstrak, hingga isi lengkap dari penelitian. Dengan jumlah publikasi yang melebihi 3 juta artikel dari 22.241 jurnal, GARUDA dapat dijadikan sumber dalam memperoleh data pendukung dalam penelitian ini yang berupa abstrak dari publikasi jurnal yang telah terindeks. Namun dalam proses identifikasi data jurnal tersebut kedalam beberapa kategori SDGs, apabila dilakukan secara manual pastinya akan membutuhkan waktu yang lama mengingat banyaknya data yang ada. Maka dari itu, diperlukan pemrosesan identifikasi data secara otomatis melalui suatu sistem guna mempercepat waktu pengerjaannya. Peluang dalam mengembangkan sistem identifikasi kategori SDGs untuk perguruan tinggi di Indonesia sangat terbuka, terlebih dari adanya kemajuan teknologi yang semakin canggih.

Identifikasi menjadi salah satu tugas NLP (*Natural Language Processing*) yang melibatkan pengkategorian teks ke dalam dua atau bahkan lebih kategori yang telah ditentukan sebelumnya. Salah satu metode yang dapat digunakan untuk identifikasi yaitu *Bidirectional Encoder Representations from Transformers* (BERT). BERT merupakan salah satu *pretrained* model dari arsitektur *transformer* yang mampu memahami konteks secara *bidirectional* (dua arah) atau lebih kompleks, sehingga membuatnya ideal untuk berbagai tugas aplikasi pemrosesan bahasa alami (NLP), termasuk di dalamnya melakukan proses identifikasi. Pada penelitian ini, menggunakan versi distilasi dari BERT yaitu DistilBERT. DistilBERT memiliki ukuran model yang lebih kecil dan waktu pelatihan yang lebih cepat dibandingkan BERT, namun tetap dapat mempertahankan performa BERT dalam melakukan berbagai tugas pemrosesan bahasa (Sanh *dkk.*, 2019).

Oleh karena itu, penggunaan DistilBERT pada penelitian ini diharapkan dapat menjadi solusi yang efisien dalam mengidentifikasi kategori SDGs berdasarkan

publikasi jurnal yang dilakukan oleh perguruan tinggi, sehingga dapat mempermudah dalam memonitor kontribusinya dalam mencapai tujuan SDGs.

1.2 Perumusan Masalah

Rumusan masalah yang diangkat pada penelitian ini adalah sebagai berikut :

1. Bagaimana mengembangkan sistem identifikasi yang dapat mengkategorikan publikasi jurnal dari perguruan tinggi ke dalam kategori *Sustainable Developments Goals* (SDGs) menggunakan model DistilBERT?
2. Bagaimana performa dari sistem identifikasi menggunakan model DistilBERT dalam mengidentifikasi kategori SDGs berdasarkan abstrak publikasi jurnal terindeks GARUDA dari perguruan tinggi?

1.3 Pembatasan Masalah

Adapun batasan masalah yang diterapkan dalam penelitian tugas akhir ini adalah sebagai berikut :

1. Hanya mengidentifikasi 3 kategori, yaitu Kehidupan Sehat dan Sejahtera (SDGs 3), Pendidikan Berkualitas (SDGs 4), serta kategori SDGs selain 3 dan 4 (*others*).
2. *Dataset* yang digunakan untuk membangun sistem identifikasi ini diperoleh dari OSDG *Community Dataset* (OSDG-CD) serta data abstrak publikasi jurnal terindeks GARUDA dari 6 Perguruan Tinggi di Indonesia.
3. Sistem hanya diimplementasikan untuk menerima input berupa abstrak dari publikasi jurnal terindeks GARUDA dari Perguruan Tinggi di Indonesia untuk dilakukan identifikasi kategori SDGs.

1.4 Tujuan

Tujuan dari penelitian ini adalah mengembangkan sistem identifikasi yang dapat memberikan kategori SDGs yang sesuai berdasarkan abstrak publikasi jurnal terindeks GARUDA dari perguruan tinggi di Indonesia menggunakan model DistilBERT.

1.5 Manfaat

Manfaat dari penilitan ini yaitu dapat mempermudah proses identifikasi kategori SDGs terhadap data publikasi jurnal terindeks GARUDA yang dimiliki perguruan tinggi di Indonesia.

1.6 Sistematika Penulisan

Dalam penyusunan laporan Tugas Akhir, menggunakan sistematika penulisan yang diuraikan sebagai berikut :

BAB I : PENDAHULUAN

Pada bab ini menguraikan tentang latar belakang yang menjadi alasan diambilnya judul, rumusan masalah yang akan diselesaikan, batasan masalah agar pembahasan dalam penelitian tidak melebar, tujuan yang ingin dicapai dari penelitian, manfaat penelitian, dan sistematika penulisan yang digunakan dalam penyusunan laporan tugas akhir.

BAB II : TINJAUAN PUSTAKA DAN DASAR TEORI

Bab ini berisikan penelitian-penelitian terdahulu yang dijadikan tinjauan dalam melakukan penelitian, serta teori-teori mengenai topik pembahasan dalam penelitian seperti sistem identifikasi dan DistilBERT.

BAB III : METODE PENELITIAN

Bab ini menjelaskan mengenai metode yang digunakan dalam penelitian, mencakup tahapan-tahapan yang dilakukan mulai dari pengumpulan data hingga proses identifikasi data.

BAB IV : HASIL PENELITIAN

Pada bab ini menyajikan hasil dari penelitian yang dilakukan, yaitu berupa model DistilBERT yang dapat mengidentifikasi kategori SDGs berdasarkan teks abstrak dari data publikasi jurnal.

BAB V : KESIMPULAN DAN SARAN

Berisikan kesimpulan yang dihasilkan dari seluruh proses penelitian, serta saran yang dapat digunakan untuk penelitian berikutnya.

BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Salah satu komponen penting dalam mengukur pencapaian *Sustainable Development Goals* (SDGs) yaitu dengan melakukan pemantauan terhadap pelaksanaannya. Dalam hal ini, adanya sistem identifikasi kategori SDGs memiliki peran krusial dalam membantu proses analisa dan evaluasi pencapaian SDGs, termasuk di dalamnya mengidentifikasi berdasarkan publikasi jurnal terindeks GARUDA.

Garba Rujukan Digital (GARUDA) menjadi salah satu *platform* referensi pengetahuan ilmiah secara nasional, yang menyediakan akses pada hasil karya ilmiah akademisi maupun peneliti di Indonesia. Pertama kali dirilis pada tahun 2018, GARUDA digunakan oleh Kemendikbud Ristek untuk mengawasi perkembangan publikasi nasional serta melakukan penilaian terhadap kualitas publikasi dengan cara akreditasi jurnal nasional (Saadah *dkk.*, 2023).

Natural Language Processing (NLP) dapat dimanfaatkan untuk mengidentifikasi dan mengklasikasikan secara otomatis teks-teks dari data publikasi yang berkaitan dengan tujuan SDGs. Keberhasilan NLP dalam melakukan tugas klasifikasi dibuktikan melalui penelitian yang dilakukan (Hanif *dkk.*, 2023). Penelitian ini bertujuan untuk mengklasifikasikan 5 bidang minat Teknik Elektro Institut Teknologi Kalimantan berdasarkan judul tugas akhir. Didapatkan hasil akurasi sebesar 86,8% dari data pengujian berupa 100 judul tugas akhir untuk seluruh bidang minat, sehingga menunjukkan bahwa NLP dapat bekerja dengan baik dalam melakukan klasifikasi.

Bidirectional Encoder Representation from Transformer (BERT) yang merupakan model berbasis *transformer*, baru-baru ini menunjukkan kinerja terbaik dalam berbagai tugas NLP (Bambroo dan Awasthi, 2021). Pada penelitian (Bagus dan Fudholi, 2021), BERT digunakan untuk melakukan klasifikasi emosi dari data teks yang diperoleh dari *twitter*. Untuk mengetahui emosi pada data yang berupa kalimat-kalimat opini dari pengguna *twitter*, dilakukan klasifikasi emosi yang

dikelompokkan menjadi sembilan kelas, yakni takut, sedih, bahagia, marah, percaya, kaget, tertarik, jijik, dan netral. Persentase tingkat akurasi model yang diperoleh dari pelatihan data sejumlah 2700 data adalah sebesar 89.1%. Penelitian ini menunjukkan BERT memiliki kemampuan yang baik dalam mengklasifikasikan teks *input* kedalam beberapa kelas emosi.

Dengan jutaan parameternya, BERT memiliki struktur kompleks dan sangat komputasional. Model DistilBERT yang lebih sederhana namun berkapasitas tinggi, digunakan sebagai solusi kendala BERT dalam penelitian ini. Dilakukan analisis sentimen pemrosesan bahasa alami (NLP) menggunakan model DistilBERT yang diterapkan pada opini teks terbuka covid-19. Penerapan DistilBERT pada penelitian ini bertujuan untuk mendeteksi perasaan positif dan negatif dari jawaban teks terbuka yang diperoleh melalui survei di masa pandemi. Dengan *dataset* pengujian didapatkan akurasi model mencapai 0,823, presisi 0,826, *recall* 0,793, dan *f1-score* 0,803, menunjukkan bahwa model tersebut dapat menggeneralisasi respon survei covid-19 dengan baik (Jojoa dkk., 2022).

Selain itu, penelitian oleh (Putri dkk., 2023) membuktikan kebutuhan komputasi DistilBERT lebih kecil dibandingkan BERT karena arsitekturnya lebih ringkas, dengan itu penggunaan memori yang dibutuhkan DistilBERT juga cenderung lebih efisien karena jumlah parameternya lebih sedikit. Penelitian ini melakukan analisis sentimen terhadap data *twitter* yang membahas mengenai pemilihan presiden Indonesia tahun 2014 dan 2019. Proses pelatihan DistilBERT pada penelitian ini memperoleh prediksi 84% lebih cepat dengan penggunaan memori GPU 79% lebih efisien dari BERT, serta nilai akurasi yang tidak berbeda jauh yakni 0,89 untuk BERT dan 0,85 untuk DistilBERT.

Penelitian serupa yang dilakukan (Fajri dkk., 2022), bertujuan membandingkan nilai akurasi BERT dan DistilBERT dalam melakukan sentimen terhadap *dataset twitter* mengenai Covid-19 yang memiliki lebih dari 2 kelas, yaitu positif, negatif, netral, sangat positif, dan sangat negatif. Berdasarkan hasil pengujian BERT mendapatkan nilai akurasi 87%, presisi 91%, *recall* 91%, dan *f1-score* sebesar 89%. Sedangkan dengan *dataset* yang sama, DistilBERT mampu menghasilkan lebih tinggi nilai akurasi sebesar 97%, presisi 99%, *recall* 99%, dan *f1-score* sebesar 99%.

Sehingga dari hasil yang didapatkan menegaskan bahwa DistilBERT lebih efektif dibandingkan BERT dalam melakukan proses klasifikasi.

Berdasarkan beberapa penelitian diatas, dapat diketahui bahwa DistilBERT mampu menghasilkan tingkat akurasi, presisi, *recall*, dan *f1-score* yang tinggi untuk berbagai tugas NLP, bahkan dapat melebihi BERT dalam beberapa kasus. Di samping itu, DistilBERT juga lebih efisien dalam waktu pemodelan dan penggunaan memori sebagai sumber daya komputasi. Maka dari itu, penggunaan model DistilBERT dalam sistem identifikasi kategori SDGs pada penelitian ini diharapkan menjadi pilihan yang tepat karena dapat memberikan keseimbangan antara efektifitas kinerja dengan efisiensi komputasi dan waktu pemodelan yang diperlukan dalam menjalankan kompleksitas tugas identifikasi.

2.2 Dasar Teori

2.2.1 *Sustainable Development Goals* (SDGs)

Sustainable Development Goals (SDGs) merupakan dokumen kesepakatan global yang bertujuan mencapai pembangunan berkelanjutan dengan memperhatikan kesetaraan tanpa ada kesenjangan, serta mengacu pada prinsip universal, integral, dan inklusif (Elvy dan Heriyanto, 2021). SDGs memiliki 17 indikator tujuan pembangunan berkelanjutan yang mencakup 4 aspek kategori: pembangunan sosial, ekonomi, lingkungan, dan kelembagaan. Tujuan SDGs nomor 3 dan 4 menjadi topik yang menarik untuk dibahas, mengingat urgensi dari kedua komponen tersebut dalam pembangunan berkelanjutan.

Memiliki jaminan kehidupan dan lingkungan yang sehat serta sejahtera merupakan keinginan seluruh manusia, hal itu sejalan dengan tujuan ketiga SDGs yaitu Kehidupan Sehat dan Sejahtera. Masalah kesehatan menjadi salah satu tantangan yang cukup sulit bagi Indonesia, karena dengan ditematkannya kesehatan sebagai salah satu substansi dalam reformulasi konsep pembangunan yang meliputi *input*, *process*, *output*, *outcome* dan *impact* bagi pembangunan serta membangun kesepahaman mengenai pentingnya pembangunan kesehatan yang saat ini harus dilakukan bersamaan untuk mencapai tujuan SDGs (Bintang *dkk.*, 2022).

Selain kesehatan, Pendidikan Berkualitas juga menjadi elemen yang penting dalam pencapaian tujuan pembangunan berkelanjutan, sehingga ditetapkanlah dalam SDGs nomor 4. Penting bagi negara untuk memastikan pemerataan pendidikan berkualitas demi memastikan warganya mendapatkan kesetaraan dan kesempatan yang sama dalam memperoleh pendidikan. Karena melalui pendidikan berkualitas yang diterima penduduk suatu bangsa dapat memberikan kontribusi dalam pembangunannya(Nurfatimah *dkk.*, 2022).

Untuk mengatasi tantangan dalam mencapai tujuan SDGs pemerintah Indonesia telah menyiapkan beberapa program seperti peningkatan kualitas pendidikan dan kesehatan, peningkatan akses dan kualitas air bersih, peningkatan akses listrik dan pengembangan energi terbarukan, serta pengurangan kemiskinan, yang harus terus di *monitoring* dalam pelaksanaannya(Sulasminingsih *dkk.*, 2024).

2.2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) merupakan salah satu bidang kecerdasan buatan (AI), yang dikembangkan untuk melakukan pemrosesan bahasa natural dalam bentuk teks maupun suara pada komputer. Bahasa natural adalah bahasa yang digunakan manusia dalam berkomunikasi satu sama lain. Sehingga, pengembangan NLP diaplikasikan pada komputer agar mampu menganalisa dan memahami bahasa sebaik manusia(Kurniawan *dkk.*, 2022).

Natural Language Processing (NLP) dapat digunakan untuk melakukan berbagai tugas terkait pemrosesan bahasa, salah satunya yaitu *text classification*. *Text classification* umumnya dilakukan untuk mempereloh hasil pengkategorian suatu teks maupun dokumen ke dalam satu atau lebih label kategori.

2.2.3 Multi-Class Classification

Text classification merupakan salah satu tugas mendasar dari NLP. Aspek penting dalam melakukan klasifikasi teks adalah jenis klasifikasinya, karena klasifikasi dapat berupa *biner*, *multi-class*, maupun *multi-label*. Dan pada penelitian ini fokus klasifikasinya menggunakan *multi-class classification*.

Multi-class classification adalah jenis tugas klasifikasi dimana setiap contoh teks atau data dikategorikan ke dalam satu dari tiga atau lebih kelas yang saling eksklusif. Dengan kata lain, setiap data hanya dapat masuk ke satu kelas dari

berbagai kelas yang tersedia. Misalnya, suatu email dapat diklasifikasikan sebagai spam, penting, atau sosial, tetapi tidak bisa masuk ke dua kelas sekaligus (Dhina dan Sumathi, 2022).

2.2.4 *Data Pre-processing*

Text preprocessing diperlukan untuk mempersiapkan data yang akan diproses agar dapat meningkatkan kualitas hasil dari suatu tugas pemrosesan data. Penjelasan berikut dapat memberikan pemahaman mengenai beberapa tahapan *pre-processing* :

1. *Data Cleaning*

Tahap awal dalam melakukan *preprocessing* adalah *data cleaning*. *Data cleaning* berfungsi untuk melakukan identifikasi terhadap data yang tidak sempurna, salah, tidak akurat atau tidak relevan. *Data cleaning* dapat dilakukan dengan menghapus karakter yang bukan angka, huruf, maupun spasi berlebihan (Joshi dan Patel, 2021).

2. *Tokenization*

Tokenisasi merupakan proses membagi teks menjadi token, yang dapat berupa kata, frasa maupun karakter. BERT sendiri memanfaatkan WordPiece Tokenizer, yang mana token pertama dari urutan token merupakan token klasifikasi khusus [CLS] dan token terakhir merupakan token pemisah [SEP] (Julianda dan Maharani, 2023).

3. *Padding dan Truncation*

Setelah melakukan tokenisasi, tahapan selanjutnya yaitu menambahkan *padding* dan *truncation* untuk memastikan semua token memiliki panjang yang sama. Token *padding* ditambahkan pada urutan token yang lebih pendek hingga mencapai panjang maksimum yang dapat diterima model, sedangkan *truncation* digunakan untuk memotong urutan token yang lebih panjang dari panjang maksimum yang dapat diterima model. Hal tersebut dilakukan karena DistilBERT memerlukan input dengan panjang yang konsisten (Husin, 2023).

4. Konversi ke Format Model

Hasil dari proses *padding* dan *truncation*, kemudian perlu dikonversi ke dalam format yang dapat diterima oleh model DistilBERT, pada penelitian ini. Proses ini dapat menggunakan *library* Transformers dari Hugging Face agar dapat menghasilkan tensor-tensor yang mampu digunakan sebagai input model.

2.2.5 *Synthetic Minority Oversampling Technique* (SMOTE)

Data yang tidak seimbang merupakan permasalahan yang umum ditemukan ketika membuat model identifikasi. Teknik yang dapat digunakan untuk mengatasi masalah ketidakseimbangan data adalah dengan melakukan *undersampling* ataupun *oversampling* data. Dengan *undersampling*, beberapa data dari kelas mayoritas akan dikurangi atau dihapus, sehingga berakibat hilangnya informasi penting yang diperlukan untuk pelatihan model. Adapun *oversampling* yang paling umum digunakan adalah *random oversampling*, yang bekerja dengan cara menduplikasi data pada kelas minoritas secara acak. Teknik ini menyebabkan adanya resiko *overfitting* karena dengan penduplikasian data maka model tidak mendapatkan informasi tambahan apapun (Erlin *dkk.*, 2022). Untuk itu, SMOTE dapat digunakan sebagai alternatif untuk mengatasi masalah tersebut.

Synthetic Minority Oversampling Technique (SMOTE) merupakan pengembangan teknik *oversampling* yang dikenalkan oleh Nithes V. Chawla. SMOTE meningkatkan jumlah distribusi data pada kelas minoritas dengan melakukan *oversampling* melalui penambahan data sintetis dari kelas minoritas. Data sintetis merupakan data baru yang dibentuk dengan pengambilan sampel data kelas minoritas dan menghubungkannya ke salah satu atau semua *k-nearest neighbors* yang dipilih secara acak. Data sintesis ini dibuat sebanyak prosentase duplikasi data yang dibutuhkan oleh kelas minoritas (Wijayanti *dkk.*, 2021). Digunakan persamaan (1) untuk mengasilkan data sintetis dengan SMOTE.

$$X_{new} = X_i + (\hat{X}_k - X_i) \times \delta \quad (1)$$

Keterangan :

X_{new} = data sintetis baru

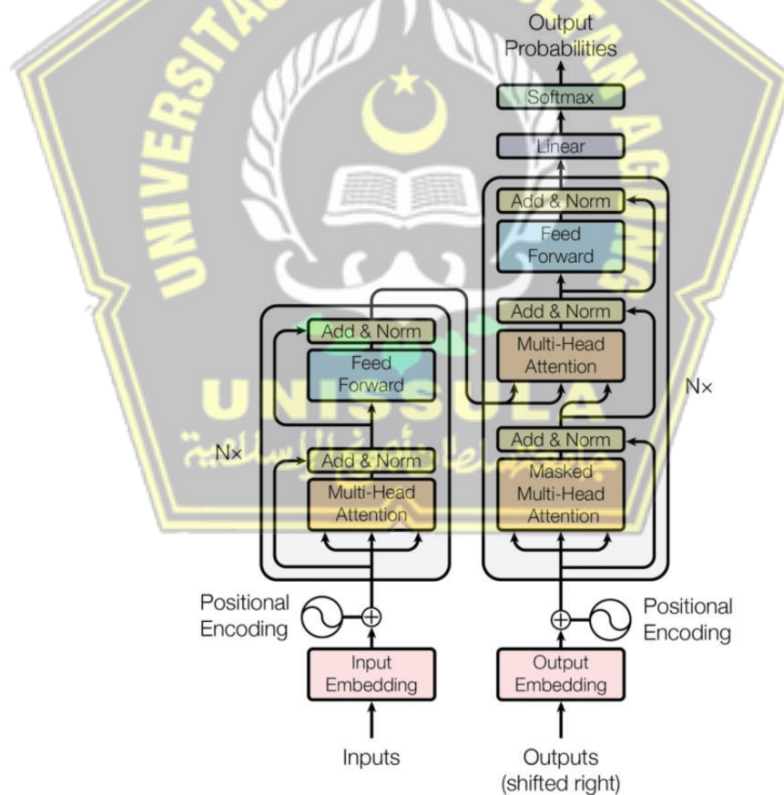
X_i = data kelas minoritas

\hat{X}_k = data dari *k-nearest neighbors* yang memiliki jarak terdekat dengan terdekat (perbedaan jarak dalam menentukan *k-nearest neighbors* dilakukan dengan menggunakan jarak *Euclidean*)

δ = bilangan acak antara 0 dan 1.

2.2.6 Transformer

Transformer merupakan model *deep learning* yang diaplikasikan pada bidang *Natural Language Processing* (NLP). *Transformer* sepenuhnya menggantungkan pada *self-attention mechanism* dalam mengkonversi pemahaman yang di peroleh dari *input*. Arsitektur transformer menggunakan bantuan *encoder* dan *decoder* dalam mengubah satu urutan ke urutan lain, serta mengandalkan *self-attention layer* dan *point-wise* dengan tujuan agar model dapat fokus pada informasi penting dari *inputannya*(Al-Faruq, 2021). Arsitektur transformer ditunjukkan pada gambar 2.1.



Gambar 2. 1 Arsitektur Transformer(Al-Faruq, 2021)

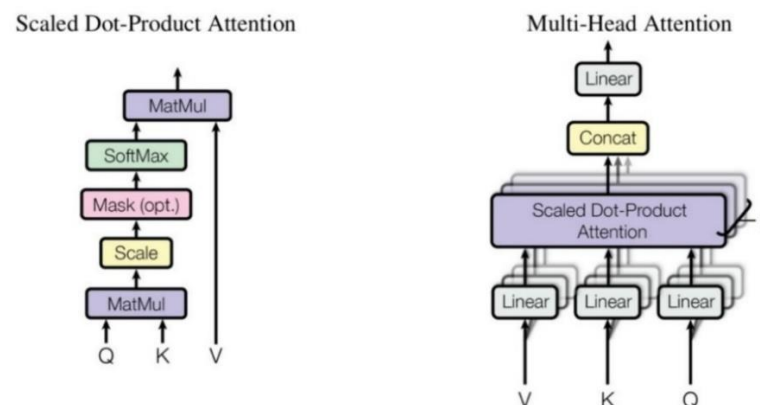
2.2.6.1. Encoder dan Decoder

Encoder berfungsi dalam membaca seluruh masukan (*input*) secara bersamaan. *Encoder* memiliki tumpukan dari $N = 6$ lapisan yang identik, setiap layernya memiliki sub-lapisan yaitu lapisan *multi-head attention* dan lapisan *feed-forward*. Serta setiap sub-lapisannya mengadopsi koneksi residual dan lapisan *normalization*.

Decoder berfungsi untuk memperoleh urutan *output* yang berupa prediksi. *Decoder* juga memiliki tumpukan dari $N=6$ lapisan yang identik, setiap lapisannya memiliki sub-lapisan yang sama dengan *encoder*, akan tetapi pada *decoder* terdapat penambahan lapisan *masked multi-head attention* di antara sub-lapisannya. Adanya lapisan *masked multi-head attention* berfungsi agar model tidak melihat (menutup) token selanjutnya saat menghasilkan token saat ini, hal ini berkaitan dengan tugas *decoder* dalam menghasilkan token demi token dalam urutan (Pratiwi dan Pardede, 2022).

2.2.6.2. Attention Mechanism

Attention berfungsi sebagai antarmuka yang menghubungkan antara *encoder* dan *decoder*. Sehingga setiap informasi penting dari *input* yang dihasilkan pada *encoder* dapat dikenali oleh *decoder* dan membuat model dapat mempelajari hubungan dari urutan *input* tersebut. *Self-attention* merupakan kategori dari *attention mechanism* yang dimanfaatkan untuk menghubungkan setiap token yang ada dalam urutan dengan token-token lain guna mendapatkan representasi lebih luas dari urutan yang sama (Al-Faruq, 2021).



Gambar 2. 2 *Scaled Dot-Product* (kiri) dan *Multi-Head Attention* (kanan) (Al-Faruq, 2021)

2.2.6.3. Scaled Dot-Product Attention

Scaled Dot-Product Attention merupakan salah satu kategori *attention mechanism* yang digunakan untuk menghitung relevansi antar token dalam urutan. Komponen penting dalam perhitungan *Scaled Dot-Product Attention* adalah *key* (K), *query* (Q), dan *value* (V) yang berupa vektor. Hasil dari *Scaled Dot-Product Attention* yaitu jumlah *value* yang dibobotkan, yang mana bobot untuk setiap *value* dipengaruhi oleh *dot-product* dari *query* dengan *keys* pada setiap token (Faruq, 2021). Setelah mendapatkan hasil *dot-product*, dilakukan proses *scaling* yaitu membaginya dengan akar kuadrat dari dimensi matriks *query-key*. *Scaling* dilakukan untuk mengantisipasi nilai yang terlalu besar sehingga mengganggu proses pelatihan. Nilai dari *scaling dot-product* selanjutnya dimasukkan pada fungsi *softmax* untuk memperoleh bobot *attention*. Bobot *attention* digunakan untuk menghitung *value* (V) yang sesuai dengan setiap token *key* (K), yang kemudian berfungsi untuk menghasilkan representasi akhir dari token dalam urutan. Matriks perhitungannya disajikan pada persamaan (2) berikut ini :

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Keterangan :

Q = *query*

K = *key*

V = *value*

d_k = *query* dan *key* dimensi

d_v = *value* dari dimensi

2.2.6.4. Multi-Head Attention

Multi-Head attention melewati *scaled dot-product attention* berkali-kali sejumlah *h* secara parallel. Tahap perhitungan *attention* ini dilakukan untuk setiap kepala (*head*) menggunakan *scaled dot-product attention* dengan matriks K, V, dan Q yang berbeda-beda. Hasil perhitungan setiap *attention* kemudian digabungkan (*concat*) lalu di proses melalui operasi pemindahan linier ke dalam dimensi yang ditentukan. Sehingga *multi-head attention* ini terdiri dari linier *layer*, *scaled dot-product attention*, *concat*, dan linier akhir (Pratiwi dan Pardede, 2022).

Untuk menggabungkan informasi dari setiap *head* digunakan persamaan (3), dimana hasil setiap *head*nya disajikan pada persamaan (4).

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (3)$$

$$\text{Wherehead}_1 = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (4)$$

2.2.6.5. Feed Forward Network

Setiap sub-lapisan *attention* yang ada didalam lapisan *encoder* dan *decoder* diberikan penambahan *Feed Forward Network* (FFN). Lapisan ini identik antara satu dengan lainnya walaupun ditambahkan dalam posisi yang terpisah. FFN terdiri dari dua lapisan linier dengan menggunakan aktivasi ReLU didalamnya. Perhitungan pada lapisan linier dan lapisan ReLU ditunjukkan pada persamaan (5)(Pratiwi dan Pardede, 2022).

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (5)$$

2.2.6.6. Embedding dan Sofmax

Untuk mengkonversi token *input* dan token *output* menjadi vektor berdimensi d_{model} , transformer menggunakan *embedding* dalam prosesnya. Selain itu, digunakan juga fungsi *softmax* untuk menghitung probabilitas dari token *output decoder* dan kemudian mengkonversinya menjadi kemungkinan token berikutnya yang akan diprediksi(Pratiwi dan Pardede, 2022).

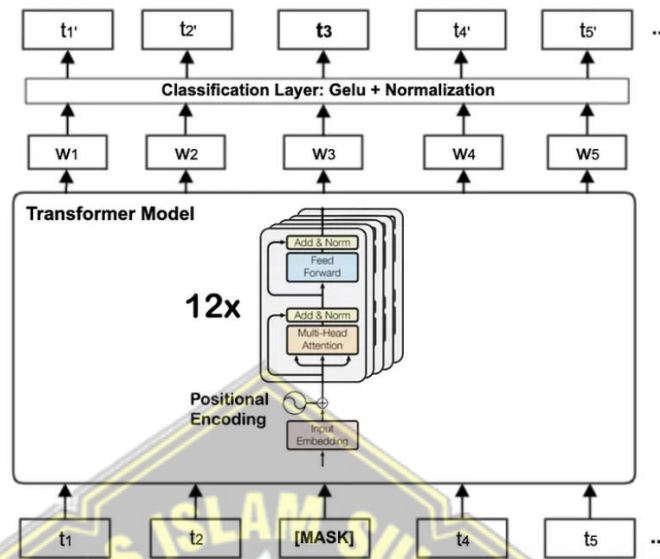
2.2.6.7. Positional Encoding

Urutan kata dalam teks *input* memberikan informasi penting terkait hubungan dan struktur kalimatnya. Untuk itu, perlu ditambahkan *positional encoding* di *embedding* token sebelum masuk ke tumpukan *encoder* dan *decoder*, guna memberikan urutan token dan informasi mengenai posisi relatif dan absolut dari token dalam urutan. *Positional encoding* mempunyai dimensi yang sama dengan *embedding* dimensi d_{model} . Fungsi sinus (6) dan fungsi cosinus (7) dengan frekuensi yang berbeda dimanfaatkan dalam menentukan urutan token pada proses *positional encoding* ini(Pratiwi dan Pardede, 2022).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (6)$$

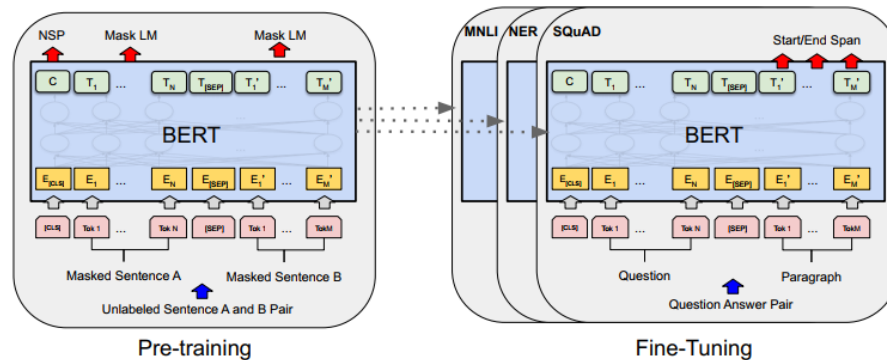
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (7)$$

2.2.7 Bidirectional Encoder Representations from Transformers (BERT)



Gambar 2. 3 Arsitektur BERT dengan 12 blok encoder

Bidirectional Encoder Representations from Transformers (BERT) merupakan model perkembangan dari arsitektur *transformer* yang hanya melakukan *encode* dalam menghasilkan sebuah model bahasa. Berbeda dengan *transformer* yang menggunakan metode terpisah, yaitu *encoder* dan *decoder* dalam membaca input dan mendapatkan hasil prediksi, BERT hanya membutuhkan *encoder* namun dalam jumlah yang lebih banyak dari aslinya. Arsitektur BERT seperti pada gambar 2. 3, memanfaatkan *transformers* dan *self-attention mechanism* dalam cara kerjanya memahami hubungan antar kata pada sebuah teks secara kontekstual. BERT termasuk dalam model yang dilatih secara *bidirectional* (dua arah) sehingga dapat membaca dan memahami teks dari kiri ke kanan, kanan ke kiri atau gabungan dari keduanya (Irfan, 2021). Arsitektur BERT memiliki 12 *layers*, 768 *hidden layer*, 12 *self-attention head*, dan total parameter sejumlah 110 juta. Dalam melakukan pemodelan bahasa, terdapat dua teknik yang terlibat dalam BERT yaitu: *pre-training* dan *fine-tuning*.



Gambar 2. 4 Pre-training dan fine tuning BERT (Devlin dkk., 2019)

2.2.7.1. Pre-Training BERT

Pre-training merupakan langkah pendekatan yang dibuat untuk melatih BERT membaca dan memahami teks bahasa beserta dengan isi konteksnya. *Pre-training* dan *fine-tuning*, secara garis besar memiliki arsitektur yang sama selain pada lapisan *outputnya*. Ketika contoh *inputan* dimasukkan ke dalam model, kemudian akan dibagi menjadi token-token yang dapat berupa kata, frasa, maupun karakter. Seperti yang ditunjukkan pada gambar 2.4, agar dapat dibaca oleh BERT, token [CLS] selalu disisipkan di setiap awal contoh *input*, dan token [SEP] ditambahkan di setiap akhir contoh *input* untuk memisahkan antar dua kalimat seperti pertanyaan/jawaban (Tandijaya dkk., 2021). Dalam *pre-training*, BERT menggunakan dua proses *unsupervised* secara bersamaan yaitu :

1. *Masked Language Model* (MLM)

Jika *transformer* menggunakan *self-attention* untuk memproses seluruh urutan token sekaligus dalam membaca atau memahami konteks input dari kedua arah (kiri dan kanan), langkah yang berbeda dilakukan BERT dengan cara menutup beberapa token input secara acak lalu memprediksi kata yang ditutup berdasarkan pada konteks token yang ada di sekitarnya. Langkah tersebut dilakukan dengan *Masked Language Model* (MLM). Sebanyak 15% dari token *input*, secara acak diubah menjadi token [MASK] untuk melatih representasi *bidirectional* lebih mendalam. Kemudian, model memprediksi kata sebenarnya dari kata yang telah ditutup [MASK] dengan memperhatikan konteks dari kata lain yang tidak ditutup [MASK] dalam urutan kata (sekitarnya). Dengan itu, MLM dapat melatih BERT untuk memahami konteks antar kata dalam satu kalimat (Irfan, 2021).

2. Next Sentence Prediction (NSP)

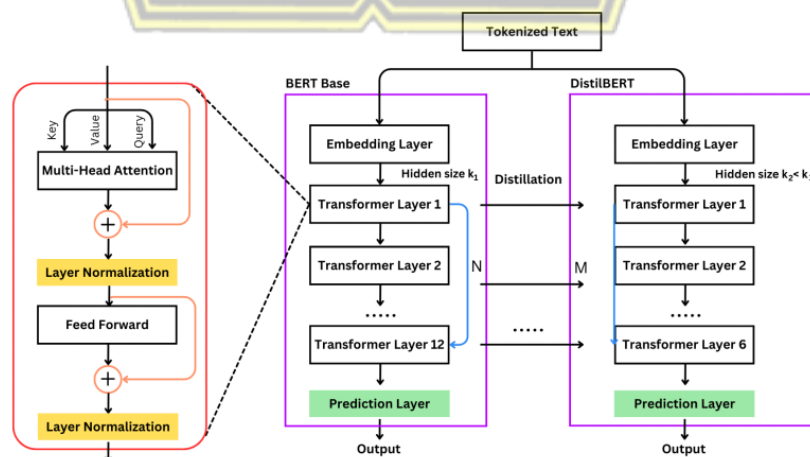
Untuk beberapa tugas, BERT perlu memahami konteks dari setiap kalimat, sehingga NSP dilakukan pada saat *pre-training*. *Next Sentence Prediction* (NSP) merupakan proses dari *pre-training* yang dapat melatih BERT dalam memahami konteks antar kalimat yang berbeda. Hasil dari NSP dapat berupa 50% probabilitas dari kalimat kedua merupakan lanjutan dari kalimat pertama dengan label IsNext atau NotNext(Irfan, 2021).

2.2.7.2. Fine-Tuning BERT

Fine-tuning melibatkan proses pelatihan ulang model yang telah di *pre-training*, dengan *task* yang lebih spesifik untuk membuat penyesuaian konfigurasi dan parameter. *Fine-tuning* sangat dibutuhkan untuk *pre-trained* model seperti BERT, karena memungkinkan BERT untuk menyesuaikan kembali dengan tugas tertentu tanpa memakan waktu dan *resource* yang lebih lama(Irfan, 2021).

2.2.7.3. DistilBERT

DistilBERT diusulkan sebagai versi distilasi dari BERT, untuk mengurangi ukuran serta meningkatkan kecepatan pelatihan model BERT. Dengan menggunakan teknik distilasi pengetahuan dari BERT selama tahap *pre-training*, DistilBERT mampu menghasilkan ukuran 40% lebih kecil, memberikan waktu inferensi 60% lebih cepat, dengan tetap mempertahankan 97% kemampuan pemahaman bahasa dari BERT(Sanh *dkk.*, 2019). Arsitektur dari DistilBERT ditunjukkan pada gambar 2. 5.



Gambar 2. 5 Arsitektur dan Komponen DistilBERT(Putri *dkk.*, 2023)

Dalam proses pelatihannya, DistilBERT menggabungkan informasi yang dimiliki oleh beberapa lapisan BERT sehingga mampu menyederhanakannya sebanyak 50% atau menjadi 6 lapisan. Bobot dari setiap lapisan BERT disesuaikan dengan lapisan DistilBERT yang jumlahnya lebih sedikit, sehingga mampu memperkecil parameter BERT *base* yang semulanya 110 juta menjadi 66 juta. Hal tersebut menjadikan DistilBERT memiliki model yang lebih sederhana sehingga mampu mempercepat waktu komputasinya (Putri *dkk.*, 2023).

2.2.8 Evaluasi Performa Model

Setelah sebuah model berhasil dirancang dan sudah di implementasikan, maka langkah selanjutnya adalah melakukan evaluasi untuk mengukur performa DistilBERT dalam mengklasifikasikan data. *Confusion matrix* merupakan salah satu metode yang digunakan untuk mengukur performa suatu model klasifikasi, dengan mendapatkan nilai *accuracy*, *presicion*, *recall* dan *f1-score* (Fajri *dkk.*, 2022). Terdapat beberapa istilah yang digunakan dalam perhitungan *confusion matrix*, di antaranya sebagai berikut :

1. *True Negative* (TN), yaitu jumlah prediksi data negatif yang diklasifikasikan dengan benar oleh sistem.
2. *True Positive* (TP), yaitu jumlah prediksi data positif yang diklasifikasikan dengan benar oleh sistem.
3. *False Positive* (FP), yaitu jumlah prediksi data negatif yang salah diklasifikasikan oleh sistem menjadi data positif.
4. *False Negative* (FN), yaitu jumlah prediksi data positif yang salah diklasifikasikan oleh sistem menjadi data negatif.

Untuk mendapatkan nilai *accuracy*, *presicion*, *recall* dan *f1-score* digunakan persamaan (8), (9), (10), dan (11) seperti berikut.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

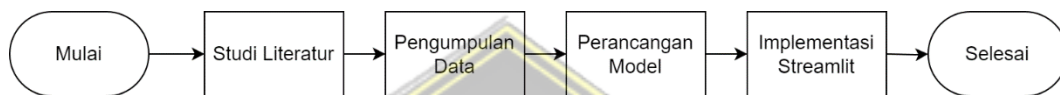
$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1 \text{ score} = \frac{2 \times (precision \times recall)}{precision + recall} \quad (11)$$

BAB III METODE PENELITIAN

3.1 Tahapan Penelitian

Pada penelitian ini menggunakan model DistilBERT untuk menghasilkan keluaran yang dapat melakukan identifikasi abstrak jurnal kedalam tiga kategori yaitu SDGs 3, SDGs 4, dan kategori *others*. Adapun tahapan yang dilakukan dalam penelitian ini digambarkan dengan *flowchart* pada gambar 3.1.



Gambar 3. 1 *Flowchart* metode penelitian

3.1.1 Studi Literatur

Untuk mendapatkan pemahaman mendalam mengenai teori-teori dalam penelitian ini, baik itu teori *Bidirectional Encoder Representation from Transformer* (BERT) hingga DistilBERT, dilakukan tinjauan terhadap beberapa sumber pengetahuan seperti artikel, jurnal, *website*, dan skripsi terdahulu.

3.1.2 Pengumpulan Data

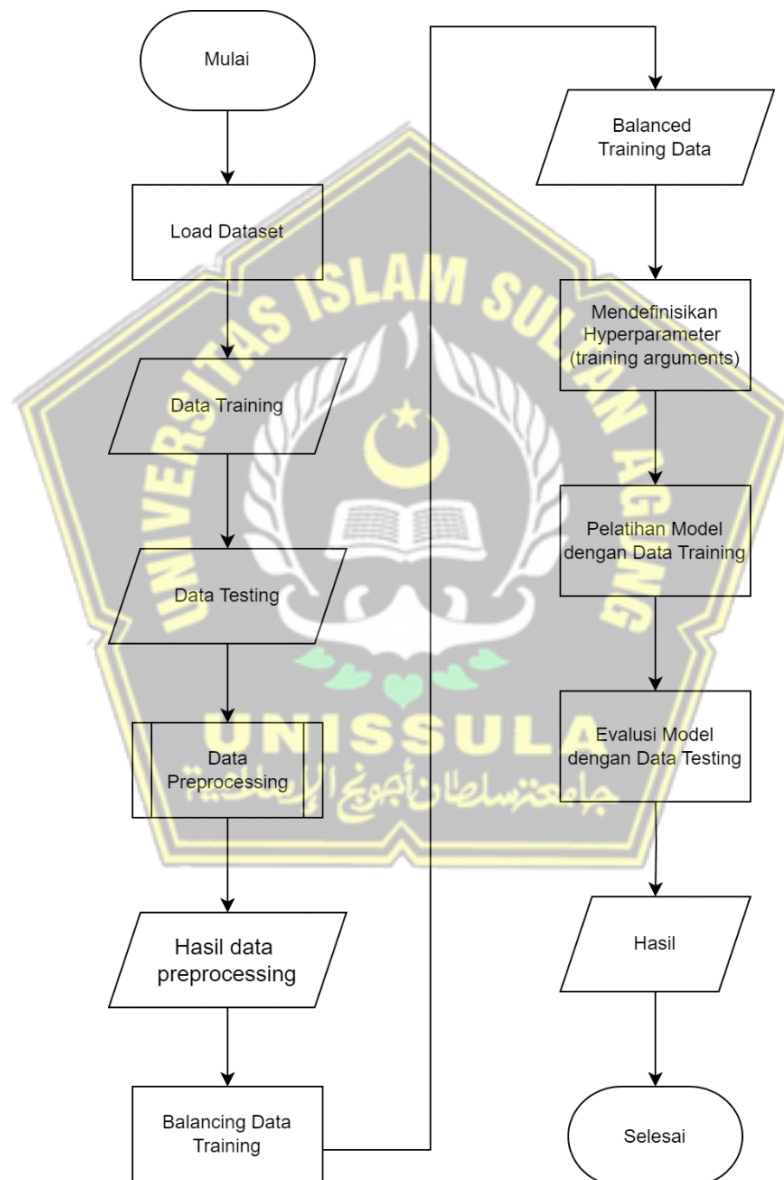
Pada penelitian ini, terdapat dua jenis *dataset* yang digunakan dalam pemodelan sistem yaitu data *training* dan data *testing*. Untuk data *training*, diambil dari data berlabel SDGs yang sudah tersedia pada OSDG *Community Dataset* (OSDG-CD)¹. Data ini berupa kumpulan paragraf dengan label kategori SDGs yang berasal dari laporan, dokumen kebijakan, dan abstrak penelitian dari sumber yang terkait dengan PBB seperti *SDG library*. Dikarenakan seluruh datanya berbahasa Inggris maka dilakukan penerjemahan data untuk memperoleh data berbahasa Indonesia dengan menggunakan *function google translate* yang ada di *google sheet*. Adapun untuk data *testing* yang digunakan untuk pengujian model diambil dari data publikasi jurnal terindeks GARUDA yang dilakukan oleh 6 perguruan tinggi di Indonesia. Data ini berupa *article_id*, asal perguruan tinggi, serta abstrak penelitian yang kemudian diberikan label kategori SDGs yang sesuai.

¹ <https://zenodo.org/records/10579179>

3.1.3 Perancangan Model

3.1.3.1. Pelatihan Model

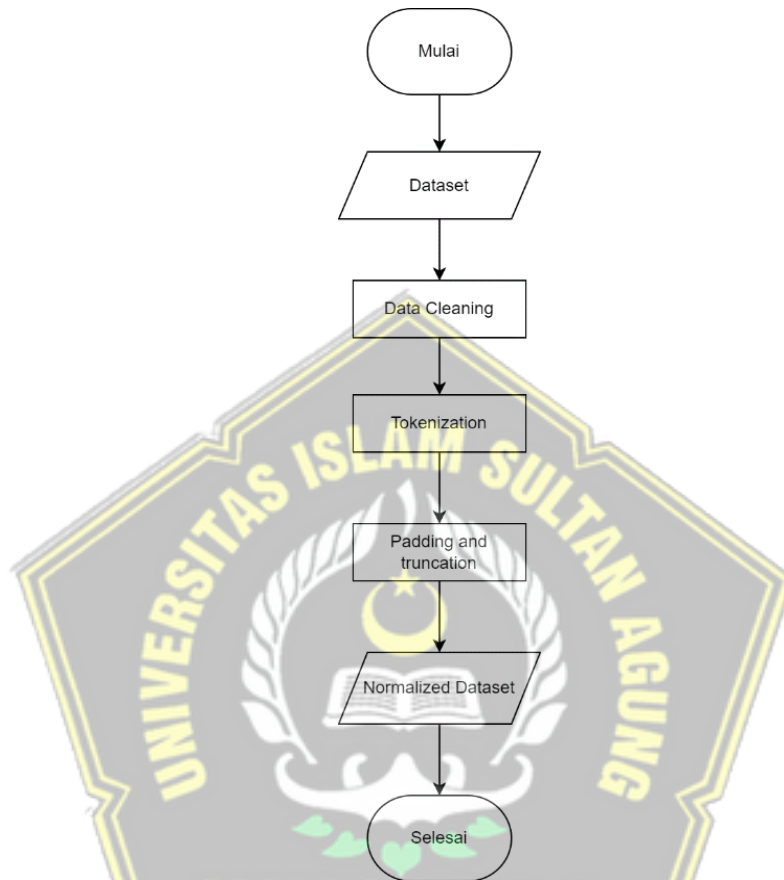
Dilakukan pelatihan secara khusus terhadap model DistilBERT yang digunakan pada penelitian ini, untuk menyesuaikannya dengan tugas identifikasi kategori SDGs yang diinginkan. Alur pelatihan model ini digambarkan melalui *flowchart* 3. 2.



Gambar 3. 2 *Flowchart* alur pemodelan sistem

Proses pelatihan model yang ditunjukkan pada gambar 3.2 secara umum terbagi menjadi 4 tahap berikut :

1. *Data pre-processing*



Gambar 3. 3 *Flowchart data pre-processing*

Sebelum *dataset* digunakan untuk pelatihan model, perlu dilakukan data *preprocessing* agar data menjadi lebih terstruktur dengan langkah-langkah seperti pada gambar 3.3. Berikut merupakan penjelasan setiap tahap *pre-processing* yang dilakukan :

- a. Langkah pertama yang dilakukan adalah memuat dataset yang akan digunakan, lalu *preprocessing* data dengan melakukan *data cleaning* untuk menghilangkan karakter selain huruf dan angka, serta spasi, kemudian mengkonversi semua hurufnya menjadi huruf kecil.
- b. Kemudian dilakukan tokenisasi, yaitu membagi teks menjadi bentuk per kata yang disebut dengan token. Misalnya kalimat seperti ‘model untuk sistem identifikasi’ akan dibagi menjadi list token [‘model’, ‘untuk’,

‘sistem’, ‘identifikasi’]. Tokenisasi ini dilakukan untuk mengkonversi data yang berupa teks menjadi bentuk numerik, menggunakan *tokenizer* DistilBERT, agar data dapat digunakan untuk pemrosesan selanjutnya.

- c. Dalam proses tokenisasi juga ditambahkan *padding* dan *truncation* untuk memastikan data memiliki panjang yang sama. Token *padding* akan ditambahkan apabila panjang data kurang dari panjang maksimum yang ditentukan atau panjang maksimum yang dapat diterima model. Sementara itu, jika data terlalu panjang akan dilakukan pemotongan (*truncation*) hingga mencapai panjang yang diinginkan.

2. *Balancing Data Training*

Tabel 3. 1 Distribusi antar kelas *data training*

No	Kelas atau kategori	Jumlah data
1	SDGs 3 (Kehidupan Sehat dan Sejahtera)	5378
2	SDGs 4 (Pendidikan Berkualitas)	7480
3	<i>Others</i>	4200

Dikarenakan terdapat ketidakseimbangan data antar kelas pada *data training*, maka perlu dilakukan *balancing data*. Salah satu cara untuk mengatasi data yang tidak seimbang adalah dengan melakukan *over-sampling* terhadap kelas minoritas agar seimbang dengan kelas mayoritas. Pada penelitian ini, menggunakan *library* “imblearn” dengan algoritma *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi *imbalance* data yang ada. Algoritma ini digunakan untuk meningkatkan jumlah sampel data di kelas minoritas, dengan cara memperkenalkan sampel-sampel baru (sintetis) yang dibuat dengan mengambil setiap sampel data di kelas minoritas yang saling berhubungan dengan beberapa sampel terdekatnya.

3. Pelatihan model

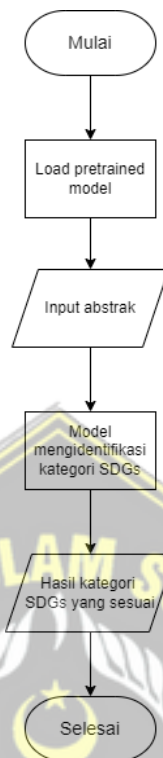
Data yang sudah diseimbangkan akan dilatih menggunakan *pretrained* model DistilBERT. Proses pelatihan model diawali dengan membuat *Data Loader* untuk memudahkan pemrosesan data, kemudian memuat model DistilBERT yang digunakan untuk identifikasi, lalu melakukan inisialisasi parameter pelatihan atau *hyperparameter* tuning dengan memanfaatkan teknik Grid

Search. Teknik ini digunakan untuk mengoptimalkan *hyperparameter* dengan menemukan konfigurasi *hyperparameter* yang menunjukkan kinerja terbaik dalam melakukan tugas identifikasi terhadap model. Hyperparameter yang coba dikombinasikan dengan teknik Grid Search adalah *learning rate*, *batch*, dan *weight decay*. *Learning rate* merupakan laju pembelajaran yang menunjukkan banyaknya informasi yang akan didapatkan model, *batch* merupakan jumlah data yang akan diproses dalam satu iterasi, dan *weight decay* adalah teknik regulasi yang dapat mengurangi *overfitting* dengan menjaga bobot tetap kecil.

4. Evaluasi model

Untuk memastikan kinerja dari model yang dikembangkan, perlu dilakukan pengujian model dengan menggunakan *data testing*. Hasil dari pengujian ini dapat dijadikan bahan untuk analisa dan evaluasi model, yang dapat dilihat melalui nilai *accuracy*, *precision*, *recall*, dan *f1-score*. *Accuracy* merupakan matriks evaluasi yang berisi nilai seberapa baik model dalam mengklasifikasikan data dengan tepat dari seluruh data uji. *Precision* menggambarkan seberapa sering model memberikan prediksi positif yang benar dari seluruh prediksi positif. *Recall* digunakan untuk mengukur seberapa banyak kelas positif yang diprediksi dengan benar. Dan *f1-score* merupakan perbandingan rata-rata dari *precision* dan *recall* yang menggambarkan kinerja model secara keseluruhan.

3.1.3.2. Inferensi Model



Gambar 3. 4 *Flowchart* inferensi model

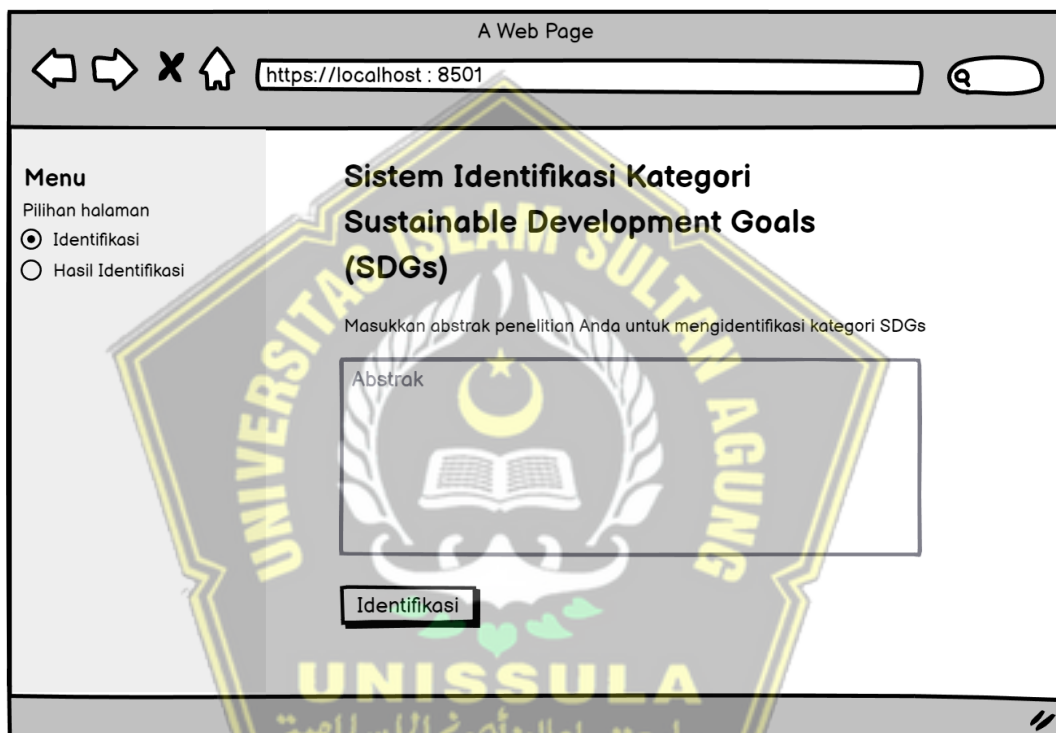
Setelah model berhasil dilatih dengan *data training*, langkah selanjutnya adalah melakukan inferensi model yang bertujuan untuk menerapkan model dalam melakukan identifikasi kategori SDGs terhadap data abstrak. *Flowchart* pada gambar 3.4 menunjukkan alur inferensi model yang dilakukan, dengan penjelasan sebagai berikut :

- a. Pertama, memuat model yang telah dilatih dengan *data training* sehingga sudah memahami pola datanya dan siap digunakan untuk mengidentifikasi data baru.
- b. Kemudian menginputkan data abstrak yang akan diidentifikasi. Data ini merupakan data baru dan tidak digunakan selama proses pelatihan model atau yang biasa disebut sebagai *data testing*.
- c. Selanjutnya model akan mengidentifikasi kategori SDGs yang sesuai untuk data abstrak yang diinputkan.
- d. Setelah itu, model akan menampilkan *output* berupa kategori SDGs, baik itu SDGs 3 (Kehidupan Sehat dan Sejahtera), SDGs 4 (Pendidikan Berkualitas), ataupun kategori *others*.

3.1.4 Implementasi Streamlit

Streamlit merupakan *framework open source* berbasis python yang dapat digunakan untuk merepresentasikan model yang telah dilatih agar mempermudah proses inferensi. Untuk itu, setelah melakukan *training* dan evaluasi pada model, langkah selanjutnya yang dilakukan adalah mengintegrasikan model ke dalam Streamlit. Rancangan tampilan streamlit yang akan dibuat adalah seperti berikut :

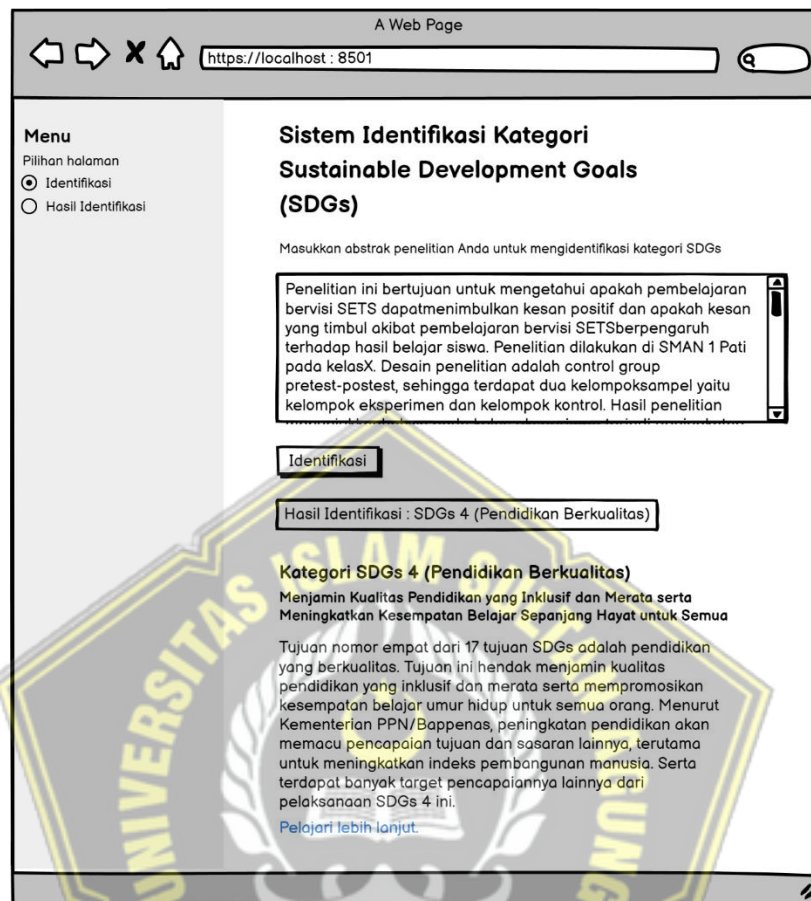
1. Tampilan Awal



Gambar 3. 5 Rancangan halaman awal sistem identifikasi

Gambar 3. 5 merupakan rancangan *interface* untuk tampilan awal sistem identifikasi. Pada halaman ini terdapat judul dan *textarea* yang digunakan untuk memasukkan abstrak penelitian sebagai input kepada sistem. Setelah itu, untuk mengidentifikasi abstrak yang dimasukkan, *user* bisa klik tombol 'Identifikasi'. Selain itu ada *side bar* yang berisi pilihan halaman yang dapat diakses, dimana pada sistem ini terdapat dua halaman yaitu halaman identifikasi dan halaman hasil identifikasi.

2. Halaman Identifikasi



Gambar 3. 6 Rancangan halaman identifikasi

Gambar 3. 6 menunjukkan halaman identifikasi yang menampilkan proses identifikasi dari teks abstrak yang dimasukkan. Akan ditampilkan hasil identifikasi kategori SDGs dari abstrak yang tersebut dan terdapat penjelasan singkat untuk setiap kategori hasilnya. Selanjutnya data hasil identifikasi ini akan masuk ke halaman hasil identifikasi.

3. Halaman Hasil Identifikasi

The screenshot shows a web browser window with the URL `https://localhost:8501`. The page title is "A Web Page". On the left, there is a "Menu" section with three options: "Pilihan halaman", "Identifikasi", and "Hasil Identifikasi" (which is selected). The main content area is titled "Hasil Identifikasi" and contains two tables.

Hasil Identifikasi

	article_id	Kategori SDGs	Perguruan Tinggi
0	2560271	SDGs 3 (Pendidikan Berkualitas)	Universitas Islam Sultan Agung
1	814757	Kategori Others	Universitas Muhammadiyah Surakarta
2	2960739	SDGs 3 (Pendidikan Berkualitas)	Universitas Islam Sultan Agung
3	2961164	Kategori Others	Universitas Kristen Satya Wacana

Hasil Identifikasi Kategori SDGs dari Perguruan Tinggi

Perguruan Tinggi	Kategori Others	SDGs 3 (Pe Berkualitas)
Universitas Islam Sultan Agung	0	1
Universitas Muhammadiyah Surakarta	1	0
Universitas Islam Sultan Agung	0	1
Universitas Kristen Satya Wacana	1	0

Gambar 3. 7 Halaman Hasil Identifikasi

Setelah berhasil mengidentifikasi kategori SDGs dari abstrak, maka hasilnya akan ditampilkan seperti pada gambar 3. 7. Pada tabel pertama ditampilkan hasil untuk setiap identifikasi abstrak secara berurutan, yang juga ditambahkan `article_id` serta asal perguruan tinggi dari abstrak tersebut. Kemudian dari hasil tersebut maka akan diketahui jumlah identifikasi setiap kategori SDGs dari perguruan tinggi dan hasilnya ditampilkan pada tabel kedua.

BAB IV

HASIL DAN ANALISIS PENELITIAN

4.1 Hasil Perancangan Model

4.1.1 Data Preprocessing

4.1.1.1. Deskripsi Dataset

Pada penelitian ini, menggunakan dua jenis *dataset* yaitu data *training* dan data *testing* yang diperoleh dari sumber berbeda. Untuk data *training*, diperoleh dari data OSDG *Community Dataset* (OSDG-CD), sedangkan untuk data *testing* diperoleh dari data publikasi jurnal terindeks GARUDA yang dilakukan oleh 6 perguruan tinggi di Indonesia. *Dataset* yang digunakan merupakan *multiclass dataset* yang dimana memiliki kategori SDGs yang digolongkan kedalam 3 kelas.

```
Info for Train:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17058 entries, 0 to 17057
Data columns (total 2 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   abstract    17058 non-null  object
1   sdgs        17058 non-null  int64
dtypes: int64(1), object(1)
memory usage: 266.7+ KB

Info for Test:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 493 entries, 0 to 492
Data columns (total 4 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   abstract        493 non-null   object
1   sdgs            493 non-null   int64
2   afiliasi_name   493 non-null   object
3   article_id     493 non-null   int64
dtypes: int64(2), object(2)
memory usage: 15.5+ KB
```

Gambar 4. 1 Informasi *dataset*

Gambar 4. 1 menampilkan informasi mengenai *dataset* seperti nama kolom, tipe data, dan jumlah nilai yang tidak null. Kedua *dataset* yang digunakan, memiliki kolom 'sdgs' dan kolom 'abstract'. Kolom 'sdgs' berisikan label kategori SDGs yang terdiri dari 3 kelas, yaitu kategori SDGs 3 diberikan label 0, SDGs 4 diberikan label 1, dan kategori *others* diberikan label 2. Adapun setiap baris dari kolom 'abstract' berisikan satu deskripsi atau abstrak suatu penelitian dan dapat memiliki 1 label dari kolom 'sdgs'. Di lain sisi, pada data *testing* terdapat dua kolom tambahan yakni

kolom “article_id” dan “afiliasi_name” yang berisikan article_id dari setiap data abstrak dan asal perguruan tinggi atau afiliasinya. Kedua kolom tersebut digunakan ketika implementasi model untuk identifikasi, namun ketika pelatihan model hanya menggunakan data pada kolom ‘abstract’ dan kolom ‘sdgs’.

Berdasarkan gambar 4. 1 dapat diketahui pula nilai yang tidak null yang menunjukkan banyaknya data, pada data *training* berjumlah 17058 sementara untuk data *testing* berjumlah 493. Jumlah data tersebut memiliki distribusi yang berbeda untuk setiap kelasnya.

Tabel 4. 1 Distribusi data untuk setiap kelas

Label kategori	Data training	Data testing
0	5378	103
1	7480	159
2	4200	231

Tabel 4. 1 menunjukkan distribusi data untuk setiap kelas kategori pada data *training* dan *testing*. Pada data *training* distribusi data untuk kategori SDGs 3(label 0) berjumlah 5378, SDGs 4(label 1) berjumlah 7480, dan kategori *others*(label 2) berjumlah 4200. Adapun pada data *testing* distribusi data untuk kategori SDGs 3 berjumlah 103, SDGs 4 berjumlah 159, dan kategori *others* berjumlah 231.

```

abstract sdgs
0 The average figure also masks large difference... 0
1 The extent to which they are akin to corruptio... 0
2 A region reporting a higher rate will not earn... 0
3 For those individuals, out-of-pocket expenses ... 0
4 In the last decade, and particularly since 201... 0
... ..
17053 Peristiwa yang terjadi belakangan ini tidak se... 2
17054 Keterlibatan dalam perbedaan dalam pembangunan... 2
17055 Artikel ini memetakan asal mula konseptual kej... 2
17056 Korupsi yang dilakukan pemerintah Meksiko meng... 2
17057 Pada skala internasional, administrasi publik ... 2

[17058 rows x 2 columns]

abstract sdgs
0 This research is aimed to explore the organiz... 1
1 Kemajuan teknologi berdampak pada kemajuan tek... 0
2 ABSTRACT : Soil is one of microorganism habita... 2
3 Penelitian ini bertujuan untuk mengetahui apak... 1
4 Penelitian ini bersifat ex post facto, artinya... 1
... ..
488 James A. Banks, profesor kulit hitam pertama y... 2
489 Collaboration is one of the 21st-century skill... 1
490 Kegiatan pengabdian masyarakat yang mengambil ... 0
491 Kota Surakarta merupakan satu diantara kota te... 2
492 Paying attention to the gap between the fenome... 2

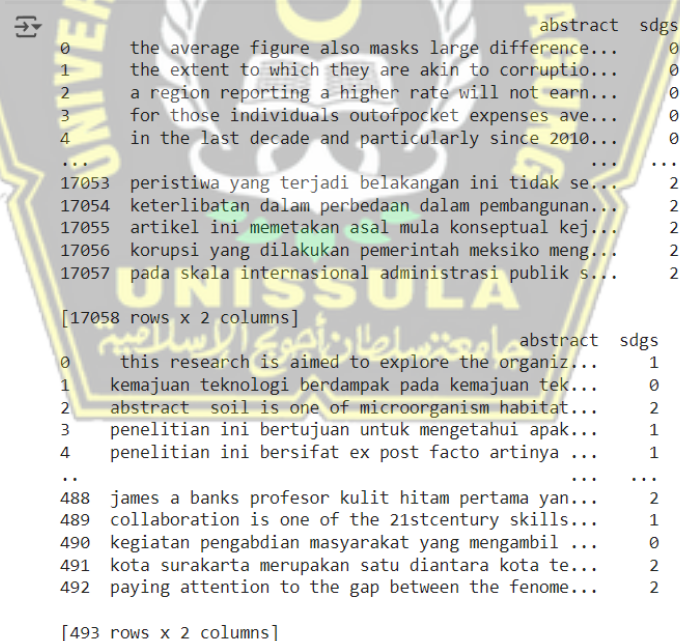
[493 rows x 2 columns]
```

Gambar 4. 2 Sampel *dataset*

Gambar 4. 2 memperlihatkan sampel *dataset* untuk memberikan wawasan mengenai struktur dan fitur *dataset* yang digunakan dalam pelatihan model pada penelitian ini. Fitur pada *dataset* yang digunakan berada pada kolom ‘abstract’ yang berisikan abstrak penelitian dan kolom ‘sdgs’ yang berisikan label kategori sdgs untuk setiap abstrak.

4.1.1.2. Data Cleaning

Proses *data cleaning* yang dilakukan adalah membersihkan data yang akan dijadikan *input* model yaitu data pada kolom *abstract* dengan menghapus karakter selain huruf, angka, serta spasi. Selain itu, diterapkan juga *lowercase* untuk mengkonversi semua huruf dalam setiap baris di kolom *abstract* menjadi huruf kecil. Pembersihan data ini berfungsi untuk memberikan data yang konsisten untuk pelatihan model, sehingga model dapat memahami pola teks dengan baik serta dapat meningkatkan akurasi dan kinerjanya. Pada proses ini, memberikan hasil pada *dataset* menjadi seperti gambar 4. 3.



```

⇒ abstract sdgs
0 the average figure also masks large difference... 0
1 the extent to which they are akin to corruptio... 0
2 a region reporting a higher rate will not earn... 0
3 for those individuals outofpocket expenses ave... 0
4 in the last decade and particularly since 2010... 0
... ..
17053 peristiwa yang terjadi belakangan ini tidak se... 2
17054 keterlibatan dalam perbedaan dalam pembangunan... 2
17055 artikel ini memetakan asal mula konseptual kej... 2
17056 korupsi yang dilakukan pemerintah meksiko meng... 2
17057 pada skala internasional administrasi publik s... 2

[17058 rows x 2 columns]
abstract sdgs
0 this research is aimed to explore the organiz... 1
1 kemajuan teknologi berdampak pada kemajuan tek... 0
2 abstract soil is one of microorganism habitat... 2
3 penelitian ini bertujuan untuk mengetahui apak... 1
4 penelitian ini bersifat ex post facto artinya ... 1
... ..
488 james a banks profesor kulit hitam pertama yan... 2
489 collaboration is one of the 21stcentury skills... 1
490 kegiatan pengabdian masyarakat yang mengambil ... 0
491 kota surakarta merupakan satu diantara kota te... 2
492 paying attention to the gap between the fenome... 2

[493 rows x 2 columns]

```

Gambar 4. 3 Hasil *data cleaning*

Untuk mengetahui perbandingan data sebelum dan sesudah dilakukan *data cleaning*, ditampilkan satu sampel hasilnya pada tabel 4. 2.

Tabel 4. 2 Perbandingan sebelum dan sesudah *data cleaning*

Sebelum	Sesudah
<p>Given the rise in chronic diseases like diabetes, an approach stressing prevention to address changing risk factors for health will be a key to helping reduce relatively more expensive hospital admissions in the future. Ideally, patients should have a trusted advisor to help them navigate through the complex number of available services. The purpose of gatekeeping is to strengthen the relationship between primary care providers and patients, thereby enhancing patientsâ€™ agency in selecting the most appropriate form of care.</p>	<p>given the rise in chronic diseases like diabetes an approach stressing prevention to address changing risk factors for health will be a key to helping reduce relatively more expensive hospital admissions in the future ideally patients should have a trusted advisor to help them navigate through the complex number of available services the purpose of gatekeeping is to strengthen the relationship between primary care providers and patients thereby enhancing patients agency in selecting the most appropriate form of care</p>

Berdasarkan tabel 4. 2 dapat diketahui bahwa proses *data cleaning* yang digunakan berhasil membersihkan teks abstrak dengan penghapusan tanda baca serta karakter yang tidak diperlukan seperti “â€™”. Kemudian, semua huruf pada teks tersebut juga dikonversi menjadi huruf kecil. Setelah dilakukan *data cleaning*, tahapan *preprocessing* selanjutnya adalah tokenisasi.

4.1.1.3. Tokenisasi

Tokenisasi merupakan proses pemecahan data yang berupa teks menjadi bentuk token. Sebelum proses tokenisasi, dilakukan pemisahan fitur (*input*) dan label dari *dataset*. Dibuat variabel ‘X’ untuk menyimpan setiap nilai dari kolom *abstract* yang kemudian nilai ini akan dijadikan *inputan* untuk model. Di sisi lain, variabel ‘y’ dibuat untuk menyimpan nilai label dari kolom *sdgs* yang akan dijadikan target untuk *output* yang diinginkan oleh model. Kemudian, dilakukan

tokenisasi terhadap teks pada variabel ‘X’ menggunakan tokenizer DistilBERT, untuk mentransformasi teks kedalam bentuk token yang dapat diproses oleh model.

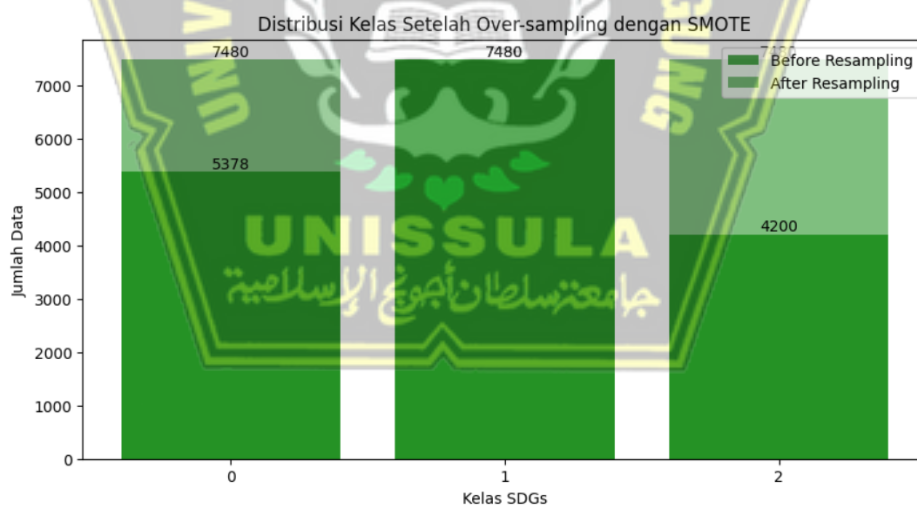
Pada proses tokenisasi ini diterapkan juga *padding* dan *truncation* untuk memastikan teks *input* yang digunakan dalam satu kali proses (*batch*) memiliki panjang yang sama, ketika teks lebih pendek maka akan ditambahkan token *padding* sedangkan teks yang panjangnya melebihi batas maksimum akan dipotong oleh *truncation*. Batas panjang maksimal token yang dapat dihasilkan ditentukan sebanyak 256 ($\text{max_length} = 256$). Hasil yang diperoleh dari proses tokenisasi ini berupa token ID yang merupakan representasi numerik dari token yang dihasilkan pada proses tokenisasi, serta *attention mask* yang berbentuk array biner untuk membedakan token yang merupakan padding (0) dan token asli dari teks (1). Setelah itu, hasil tokenisasi ini akan digunakan sebagai *inputan* dan diproses lebih lanjut oleh model DistilBERT yang digunakan pada penelitian ini. Pada tabel 4. 3 ditampilkan contoh hasil pemecahan data yang berupa teks menjadi token-token.

Tabel 4. 3 Proses Tokenisasi

Teks asli	Token
<p>the average figure also masks large differences across regions in kazakhstan the number of annual contacts ranges from 20 in astana to 97 in mangystau and some parts of the population are likely to have very limited access to primary care in addition poor coverage of outpatient prescription medicines limits both the effectiveness and appeal of care at phc level</p>	<p>[the] [average] [figure] [also] [masks] [large] [differences] [across] [regions] [in] [kazakhstan] [the] [number] [of] [annual] [contacts] [ranges] [from] [20] [in] [as] [##tana] [to] [97] [in] [man] [##gy] [##sta] [##u] [and] [some] [parts] [of] [the] [population] [are] [likely] [to] [have] [very] [limited] [access] [to] [primary] [care] [in] [addition] [poor] [coverage] [of] [out] [##patient] [prescription] [medicines] [limits] [both] [the] [effectiveness] [and] [appeal] [of] [care] [at] [ph] [##c] [level]</p>

4.1.2 *Balancing Data Training*

Terdapat ketidakseimbangan distribusi data untuk setiap kelas pada data *training*, seperti yang terlihat tabel 4.1. Kelas yang memiliki jumlah data banyak disebut kelas mayoritas, sedangkan kelas dengan jumlah data sedikit disebut kelas minoritas. Data yang tidak seimbang dapat berpengaruh pada kinerja model. Model yang dilatih dengan data yang tidak seimbang akan cenderung mengklasifikasikan kelas mayoritas secara berlebihan dan mengabaikan kelas minoritas, sehingga akan mengakibatkan banyak kesalahan klasifikasi. Selain itu, distribusi data yang tidak seimbang juga dapat mengakibatkan *overfitting*. *Overfitting* merupakan kondisi dimana model memiliki kinerja yang baik terhadap data *training*, namun tidak pada data *testing*. Oleh karena itu, pada penelitian ini dilakukan penyeimbangan data *training* menggunakan SMOTE sebelum data digunakan untuk pelatihan model. SMOTE melakukan oversampling untuk kelas minoritas agar jumlah datanya menjadi seimbang dengan kelas mayoritas. Hasil dari proses *balancing* ini ditunjukkan pada gambar 4. 4.



Gambar 4. 4 Jumlah data hasil *balancing*

Gambar 4. 4 memperlihatkan distribusi data *training* setelah dilakukan *resampling* menggunakan SMOTE. Warna hijau tua menunjukkan jumlah data asli sebelum di seimbangkan, dan warna hijau muda menunjukkan jumlah data setelah dilakukan *resampling* dengan SMOTE. Jumlah data kelas minoritas diseimbangkan dengan kelas mayoritas sehingga jumlah datanya menjadi 7480 untuk setiap kelas.

4.1.3 Hasil Pelatihan Model

Data *training* yang telah diseimbangkan sebelumnya, digunakan untuk melatih model DistilBERT dalam melakukan identifikasi kategori SDGs. Pada proses pelatihan model, digunakan teknik Grid Search untuk efisiensi waktu dalam melakukan eksperimen. Teknik ini digunakan untuk mendapatkan konfigurasi *hyperparameter* dengan performa terbaik terhadap model. *Hyperparameter* yang dikombinasikan dengan dengan Grid Search pada penelitian ini adalah *learning rate*, *batch training*, dan *weight decay*. Tabel 4. 4 menampilkan nilai dari setiap *hyperparameter* yang coba dikombinasikan.

Tabel 4. 4 Konfigurasi *hyperparameter* dengan grid search

Hyperparameter	Nilai
learning rate	[5e-5, 4e-5, 3e-5, 2e-5, 1e-5, 1e-4]
batch training	[16, 32]
weight decay	[0.05, 0.1, 0.08]

Dengan konfigurasi *hyperparameter* yang telah ditentukan pada tabel 4. 3, pelatihan model dilakukan selama 10 *epoch*. Dari proses pelatihan model dengan menerapkan teknik grid search, diperoleh hasil *accuracy* terbaik terhadap data training sebesar 86,73% pada konfigurasi *learning rate* 0.0001 atau 1e-4, *batch* 32, dan *weight decay* 0.08. Dari hasil yang didapatkan, menunjukkan bahwa model mampu untuk menggeneralisasi dengan baik pada data training. Selanjutnya perlu dilakukan evaluasi kinerja model terhadap data *testing* untuk memastikan bahwa model tidak *overfitting*.

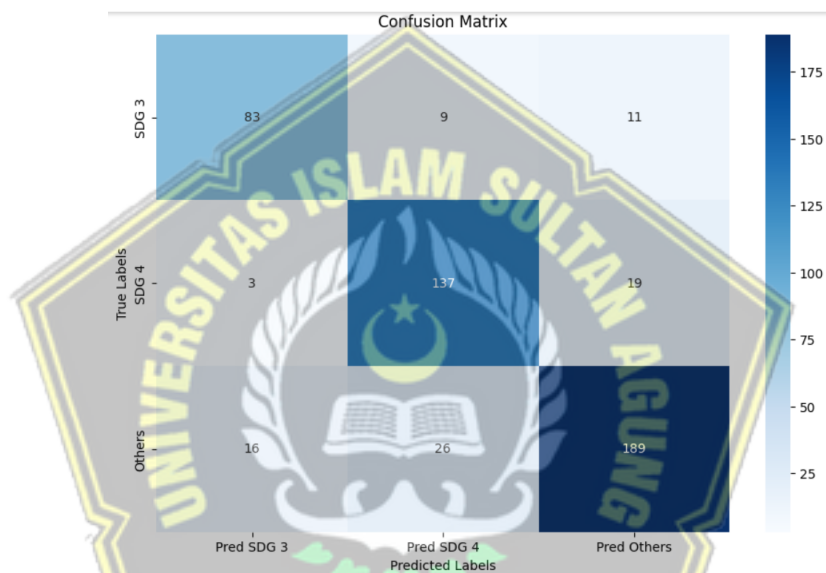
4.2 Hasil Evaluasi

Untuk memastikan kinerja dari model, dilakukan evaluasi menggunakan *confusion matrix* dengan mengukur matrik *accuracy*, *precision*, *recall*, dan *f1-score* terhadap data *training* dan data *testing*.

Tabel 4. 5 Hasil evaluasi model

Jenis dataset	Loss	Accuracy	Precision	Recall	F1-score
Data training	0.283548	86,73%	88,34%	86,73%	86,59%
Data testing	0.623751	82,56%	82,94%	82,56%	82,57%

Tabel 4. 5 menampilkan hasil evaluasi model terhadap data training dan data *testing*. Hasil *accuracy*, *precision*, *recall*, dan *f1-score* tersebut memiliki nilai yang cukup tinggi sehingga menunjukkan bahwa model memiliki kinerja yang baik dalam melakukan prediksi pada kedua dataset. Dapat diketahui pula bahwa model mengalami *overfitting*, dimana hasil kinerjanya menurun ketika digunakan pada data *testing*. Namun, *overfitting* yang terjadi masih ternilai kecil dan model masih mampu menghasilkan *accuracy* sebesar 82,56% terhadap data *testing*. Hasil tersebut diperoleh dari pemetaan nilai *confusion matrix*s seperti pada gambar 4. 5.



Gambar 4. 5 Hasil Confision Matrix

Gambar 4. 5 merupakan hasil *confusion matrix* yang memetakan jumlah prediksi benar dan salah untuk masing-masing kelas. Dari hasil tersebut dapat diketahui bahwa kategori *others* sering salah diidentifikasi menjadi kategori SDGs 4, begitu pula sebaliknya. Hal tersebut mungkin saja terjadi karena isi abstrak penelitian terkadang memiliki kemiripan karakteristik atau kata-kata kunci didalamnya, sehingga membuat model sulit untuk mengidentifikasi dengan benar.

Kemudian, untuk mengetahui kemampuan model dalam mengidentifikasi setiap kelas kategori SDGs pada data *testing* dilakukan evaluasi menggunakan metrik yang sama dan diperoleh hasil seperti pada tabel 4. 6.

Tabel 4. 6 Hasil evaluasi setiap kelas data *testing*

Kategori SDGs	Precision	Recall	F1-score
SDGs 3	0.81	0.81	0.81
SDGs 4	0.80	0.86	0.83
Others	0.86	0.82	0.84

Dari hasil evaluasi pada tabel 4. 6 dapat disimpulkan bahwa model yang dibangun memiliki performa yang cukup seimbang untuk setiap kelas kategori SDGs. Model memiliki performa terbaik dalam melakukan identifikasi kategori Others, dengan nilai *presicion* 0.86 atau dalam prosentase sebesar 86%, yang menunjukkan bahwa banyak prediksi positif yang berhasil diprediksi dengan benar oleh model.

4.3 Hasil Implementasi Steamlit

4.3.1 Tampilan Awal Sistem Identifikasi

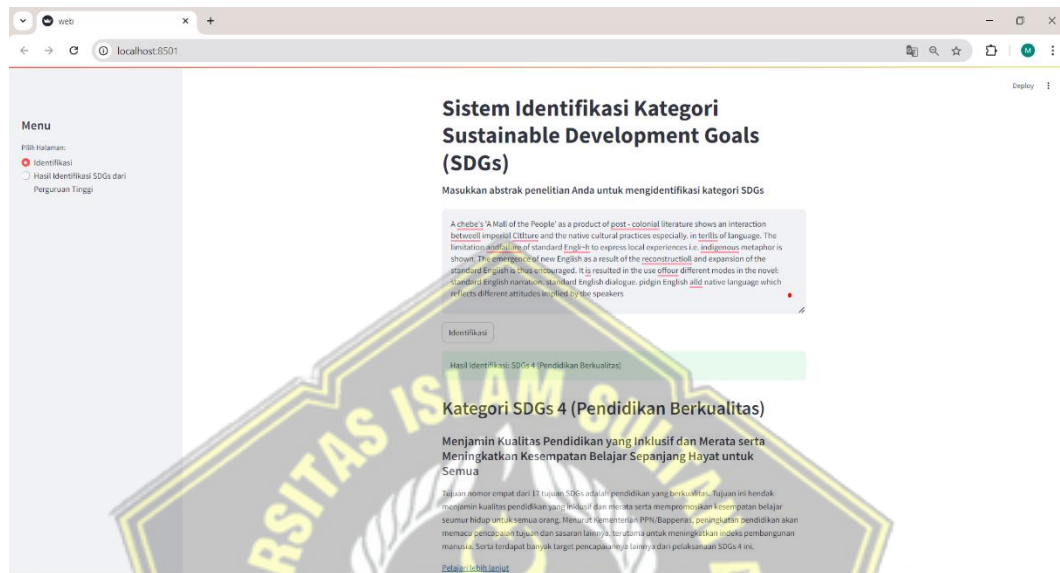


Gambar 4. 6 Tampilan awal streamlit

Model DistilBERT yang telah dilatih sebelumnya, diintegrasikan dengan *framework* Streamlit untuk memudahkan pengguna dalam melakukan identifikasi. Terlihat pada gambar 4. 6 merupakan hasil implementasi Streamlit untuk tampilan awal dari sistem identifikasi kategori SDGs, yang memiliki dua pilihan halaman yaitu halaman identifikasi dan juga halaman hasil identifikasi. Pada tampilan ini terdapat *text area* yang digunakan untuk memberikan *input* berupa abstrak penelitian yang dimiliki oleh *user*. Lalu untuk menjalankan proses identifikasi, *user*

bisa menekan tombol ‘Identifikasi’ yang ada di bawah *text area*. Setelah itu, akan keluar hasil identifikasi kategori SDGs yang relevan dengan abstrak yang *diinputkan*.

4.3.2 Halaman Hasil Sistem Identifikasi



Gambar 4. 7 Hasil identifikasi SDGs 4

Pada gambar 4. 7 merupakan tampilan dari halaman identifikasi yang digunakan untuk mengidentifikasi kategori SDGs dari abstrak yang *diinputkan*. Pada halaman ini, hasil identifikasi kategori SDGs ditampilkan pada *alert box* dengan warna hijau yang berarti bahwa sistem telah berhasil mengidentifikasi abstrak yang *diinputkan*. Kemudian, ditampilkan juga penjelasan singkat secara dinamis sesuai dengan identifikasi kategori SDGs yang dihasilkan.

Sebagai contoh untuk percobaan pertama pada gambar 4. 7 diberikan *input* berupa abstrak mengenai kurikulum yang banyak diterapkan oleh lembaga pendidikan tinggi di Indonesia. Setelah itu, sistem melakukan identifikasi kategori SDGs dan menampilkan hasilnya sebagai kategori SDGs 4 (Pendidikan Berkualitas), dibawah hasil identifikasi tersebut terdapat penjelasan untuk memberikan informasi tambahan terkait dengan SDGs 4 itu sendiri.

Sistem Identifikasi Kategori Sustainable Development Goals (SDGs)

Masukkan abstrak penelitian Anda untuk mengidentifikasi kategori SDGs

ABSTRAK Latar belakang: penderita diabetes melitus terus mengalami peningkatan tiap tahunnya. Laporan menunjukkan bahwa terapi pijat merupakan terapi komplementer yang digunakan di Indonesia serta memiliki banyak manfaat. Tujuan: menggali lebih dalam mengenai pengaruh terapi pijat terhadap tingkat kadar glukosa darah pada pasien dengan diabetes melitus. Metode: Pencarian artikel dilakukan menggunakan Science Direct, Medline, Google Search dan Pro Quest untuk menemukan artikel sesuai kriteria inklusi dan eksklusi. Artikel yang sesuai dengan kriteria yang ditetapkan penulis dianalisis, ditentukan level dari evidencenya, diekstraksi kemudian disintesis. Hasil: terapi pijat secara signifikan mampu mengontrol kadar glukosa darah pada pasien...

Identifikasi

Hasil identifikasi: SDGs 3 (Kehidupan Sehat dan Sejahtera)

Kategori SDGs 3 (Kehidupan Sehat dan Sejahtera)

Menjamin Kehidupan yang Sehat dan Meningkatkan Kesejahteraan Seluruh Penduduk Semua Usia

Tujuan nomor tiga dari 17 tujuan SDGs adalah kehidupan sehat dan sejahtera atau good health and well-being. Tujuan ini menjamin kehidupan yang sehat dan mendorong kesejahteraan bagi semua orang di segala usia. Fokus dan tujuan ini mencakup berbagai hal mulai dari menjamin gizi masyarakat, sistem kesehatan nasional, akses kesehatan dan pelayanan, ketepatan perencanaan (P5), serta sanitasi dan air bersih. Serta terdapat banyak target pencapaian yang lainnya dari pelaksanaan SDGs 3 ini.

[Pelajari lebih lanjut](#)

Gambar 4. 8 Hasil identifikasi SDGs 3

Percobaan kedua pada gambar 4. 8 menunjukkan hasil identifikasi abstrak yang dikategorikan sebagai SDGs 3 (Kehidupan Sehat dan Sejahtera). Abstrak yang diberikan berisi tentang helminthiasis yaitu jenis penyakit pada usus manusia, sehingga membuatnya relevan dengan kategori SDGs 3 yang berkaitan dengan kesehatan. Kemudian pada halaman hasil ini ditampilkan pula informasi tambahan mengenai fokus utama dari tujuan SDGs 3.

Sistem Identifikasi Kategori Sustainable Development Goals (SDGs)

Masukkan abstrak penelitian Anda untuk mengidentifikasi kategori SDGs

Abstrak Kegiatan ini bertujuan untuk 2) Memberikan pembinaan kepada mahasiswa terkait pelaksanaan kegiatan bimbingan belajar yang 2) Membekalkan pemateri/program kepada mahasiswa dalam meningkatkan kegiatan bimbingan belajar yang 3) Menentukan strategi dalam pelaksanaan kegiatan bimbingan belajar dengan mempertimbangkan metode yang dilaksanakan pada kegiatan ini adalah berdasarkan hasil diskusi antara pemateri dan siswa, selanjutnya dipaparkan bahwa prioritas masalah yang dapat dipecahkan untuk mendukung selama pelaksanaan program PKM-PIB adalah perbaikan pematerian, pendampingan dan praktik pelaksanaan serta kegiatan bimbingan belajar yang perlu diperhatikan dalam kegiatan ini adalah bagaimana pesan pembaharuan dapat disampaikan secara...

Identifikasi

Hasil identifikasi: Kategori Others

Kategori Others

Indikator SDGs yang termasuk kategori Others pada sistem ini adalah kategori selain SDGs 3 & 4. Yang mana SDGs sendiri memiliki 17 indikator kategori sebagai berikut: (1) Tanpa Kemiskinan; (2) Tanpa Kelaparan; (3) Kehidupan Sehat dan Sejahtera; (4) Pendidikan Berkualitas; (5) Kesetaraan Gender; (6) Air Bersih dan Sanitasi Layak; (7) Energi Bersih dan Terjangkau; (8) Pekerjaan Layak dan Pertumbuhan Ekonomi; (9) Industri, Inovasi dan Infrastruktur; (10) Berkualitas Kesejahteraan; (11) Kota dan Permukiman yang Berkelanjutan; (12) Konsumsi dan Produksi yang Bertanggung Jawab; (13) Penanganan Perubahan Iklim; (14) Ekosistem Lautan; (15) Ekosistem Daratan; (16) Perdamaian, Keadilan dan Kolaborasi yang Tangguh; (17) Komitmen untuk Mencapai Tujuan. Pelajari lebih lanjut mengenai 17 indikator SDGs dan target pencapaiannya.

[Pelajari lebih lanjut](#)

Gambar 4. 9 Hasil identifikasi kategori others

Pada gambar 4. 9 menunjukkan hasil identifikasi abstrak sebagai kategori *Others*. Kategori *others* pada penelitian ini merupakan pengkategorian untuk tujuan SDGs selain pada kategori SDGs 3 dan 4, yang mana SDGs sendiri memiliki 17 tujuan atau kategori. Penjelasan dibawah hasil identifikasi untuk kategori ini juga menampilkan 17 tujuan yang ada pada SDGs, selain itu terdapat tautan untuk mengetahui informasi lebih lanjut terkait tujuan-tujuan SDGs tersebut.

4.3.3 Halaman Hasil Identifikasi

Hasil Identifikasi SDGs

article_id	kategori_sdgs	perguruan_tinggi
1767262	SDGs 4 (Pendidikan Berkualitas)	Universitas Katolik Soegijapranata
2533696	Kategori Others	Universitas Muhammadiyah Surakarta
3297854	SDGs 3 (Kehidupan Sehat dan Sejahtera)	Universitas Islam Sultan Agung
1912109	Kategori Others	Universitas Islam Sultan Agung
2225846	SDGs 3 (Kehidupan Sehat dan Sejahtera)	Universitas Muhammadiyah Semarang

Hasil Identifikasi SDGs dari Perguruan Tinggi

perguruan_tinggi	Kategori Others	SDGs 3 (Kehidupan Sehat dan Sejahtera)	SDGs 4 (Pendidikan Berkualitas)
Universitas Islam Sultan Agung	1	1	0
Universitas Katolik Soegijapranata	0	0	0
Universitas Muhammadiyah Semarang	0	1	0
Universitas Muhammadiyah Surakarta	1	0	0

Gambar 4. 10 Halaman Hasil Identifikasi

Gambar 4. 10 merupakan tampilan dari halaman hasil identifikasi. Setelah abstrak berhasil diidentifikasi kategori SDGs-nya, maka akan ditampilkan hasilnya pada halaman ini. Pada tabel pertama, menampilkan hasil dari setiap identifikasi abstrak yang dilakukan dan kemudian ditampilkan juga article_id serta asal perguruan tingginya. Setelah itu, jumlah hasil identifikasi setiap kategori SDGs untuk masing-masing perguruan tinggi ditampilkan pada tabel kedua.

4.3.4 Hasil Tampilan *Error*



Gambar 4. 11 Hasil tampilan *error*

Gambar 4. 11 menunjukkan tampilan *error* pada sistem yang terjadi karena belum adanya input yang diberikan, namun *user* telah menekan tombol 'Identifikasi' sehingga muncul *alert box* berwarna merah atau *error* untuk memasukkan abstrak yang ingin diidentifikasi.

4.4 Hasil Pengujian Sistem

Pada tahap ini, digunakan metode *black box testing* untuk menguji kinerja sistem identifikasi. Metode ini berfungsi untuk memastikan bahwa sistem identifikasi bekerja sesuai dengan spesifikasi dan tujuan yang telah ditetapkan yaitu mengidentifikasi abstrak (*input*) dan memberikan *output* kategori SDGs yang relevan. Dengan metode *black box testing* dapat diukur pula akurasi dan konsistensi sistem dalam mengidentifikasi, apakah hasilnya sesuai yang diharapkan untuk berbagai data uji yang diinputkan. Untuk itu, strategi pengujiannya adalah memberikan *input* kepada sistem berupa data uji dengan kategori SDGs yang berbeda-beda untuk melihat *output* yang dihasilkan apakah sesuai dengan harapan atau tidak. Hasil dan kesimpulan untuk setiap pengujian dituliskan pada tabel 4. 7.

Tabel 4. 7 Hasil Pengujian Sistem Identifikasi

<i>Input</i>	Kategori SDGs yang diharapkan	<i>Output</i>	Kesimpulan
Defek telinga unilateral ataupun bilateral dapat disebabkan oleh berbagai faktor	SDGs 3 (Kehidupan Sehat dan Sejahtera)	SDGs 3 (Kehidupan Sehat dan Sejahtera)	Sistem mengidentifikasi dengan akurat
Pembelajaran IPA di SMPN 3 Mranggen kompetensi terkait....	SDGs 4 (Pendidikan Berkualitas)	SDGs 4 (Pendidikan Berkualitas)	Sistem mengidentifikasi dengan akurat
Indonesia is vulnerable to earthquake and tsunami disaster....	Kategori <i>Others</i>	Kategori <i>Others</i>	Sistem mengidentifikasi dengan akurat
Regulations governing the crime of corruption has changed	Kategori <i>Others</i>	Kategori <i>Others</i>	Sistem mengidentifikasi dengan akurat
Chronic obstruction pulmonary disease (COPD) is a lung disease.....	SDGs 3 (Kehidupan Sehat dan Sejahtera)	SDGs 3 (Kehidupan Sehat dan Sejahtera)	Sistem mengidentifikasi dengan akurat
This research is aimed to explore the organization	SDGs 4 (Pendidikan Berkualitas)	Kategori <i>Others</i>	Hasil identifikasi sistem tidak akurat

<i>Input</i>	Kategori SDGs yang diharapkan	<i>Output</i>	Kesimpulan
behavior of Universitas Muhammadiyah Semarang.....			
This research aims to improve critical thinking skills on thematic learning.....	SDGs 4 (Pendidikan Berkualitas)	SDGs 4 (Pendidikan Berkualitas)	Sistem mengidentifikasi dengan akurat
Software berkualitas tinggi adalah software yang tidak ditemukan cacat selama pemeriksaan.....	Kategori <i>Others</i>	Kategori <i>Others</i>	Sistem mengidentifikasi dengan akurat
.... hubungan antara dukungan sosial dengan resiliensi pada narapidana laki-laki kasus narkoba pada masa pandemi Covid-19.....	SDGs 3 (Kehidupan Sehat dan Sejahtera)	Kategori <i>Others</i>	Hasil identifikasi sistem tidak akurat
Sampling technique that has the highest	SDGs 4 (Pendidikan Berkualitas)	SDGs 4 (Pendidikan Berkualitas)	Sistem mengidentifikasi dengan akurat

<i>Input</i>	Kategori SDGs yang diharapkan	<i>Output</i>	Kesimpulan
precision in this research is the stratified random sampling technique.....			

Dapat diamati pada tabel 4. 7 menunjukkan hasil pengujian *black box* dari sepuluh data *input*, terdapat delapan data yang diidentifikasi dengan benar oleh sistem sementara dua data lainnya salah diidentifikasi. Hasil ini menunjukkan sistem memiliki kinerja yang cukup baik dalam mengidentifikasi dan memberikan kategori SDGs yang relevan terhadap data baru yang belum pernah dilihat oleh model.

Kesalahan identifikasi mungkin saja terjadi karena adanya kata-kata kunci pada abstrak yang dapat merujuk ke beberapa kategori, sehingga model sulit untuk mengenali pola kalimatnya dan salah mengidentifikasi. Sebagai contoh pada abstrak “*This research is aimed to explore the organization behavior of Universitas Muhammadiyah Semarang.....*” hasil identifikasinya menunjukkan kategori *others* bukan SDGs 4, dikarenakan di dalam abstrak tersebut fokus utamanya membahas mengenai kelembagaan di Universitas Muhammadiyah Semarang seperti kepemimpinan, norma dan tradisi, serta manajemen organisasinya, sehingga tidak teridentifikasi sebagai SDGs 4 yang fokus utamanya berkaitan dengan aspek-aspek pendidikan seperti kualitas dan hasil pembelajaran.

Dari hasil pengujian secara keseluruhan, dapat disimpulkan bahwa sistem yang dibangun mampu memberikan hasil identifikasi kategori SDGs yang akurat dari data abstrak yang diinputkan.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Pada penelitian ini, telah dibangun sistem identifikasi kategori SDGs dengan melatih model DistilBERT sehingga mampu mengkategorikan abstrak jurnal penelitian ke dalam kategori SDGs 3, SDGs 4, maupun kategori *Others*. Berdasarkan hasil pengujian yang dilakukan, model menunjukkan performa terbaik pada data *testing* dengan nilai *accuracy* 82,56%, *precision* 82,94%, *recall* 82,56%, dan *f1-score* sebesar 82,57%. Hasil ini menunjukkan bahwa model memiliki kinerja yang cukup baik dalam mengidentifikasi kategori SDGs. Dengan demikian, model DistilBERT yang dilatih telah berhasil menunjukkan kemampuan yang memadai dalam melakukan tugas identifikasi ini, sehingga dapat diandalkan untuk mendukung klasifikasi publikasi jurnal terindeks GARUDA yang dimiliki perguruan tinggi di Indonesia sesuai dengan kategori SDGs.

5.2 Saran

Meskipun pada penelitian ini telah menghasilkan performa model yang cukup baik, terdapat beberapa saran untuk peningkatan pada penelitian selanjutnya yaitu sebagai berikut :

1. Untuk meningkatkan kinerja model, disarankan menggunakan *dataset* yang berasal dari satu sumber yang sama yaitu dari jurnal terindeks GARUDA.
2. Penelitian ini hanya mengidentifikasi data abstrak ke dalam tiga kategori SDGs, untuk itu pada penelitian selanjutnya disarankan untuk memperluas cakupan identifikasi menjadi 17 kategori SDGs yang ada.
3. Disarankan untuk penelitian selanjutnya dapat dilakukan dengan *multi-label classification*, dikarenakan isi dari abstrak jurnal penelitian terkadang memiliki konteks beririsan sehingga dapat diidentifikasi ke beberapa kelas kategori.

DAFTAR PUSTAKA

- Al-Faruq, U.A. (2021) “Implementasi Arsitektur Transformer Pada Image Captioning Dengan Bahas Indonesia.” Tersedia pada: <https://dspace.uui.ac.id/handle/123456789/36130>.
- Bagus, A.T. dan Fudholi, D.H. (2021) “Klasifikasi Emosi pada Teks Dengan Menggunakan Metode Deep Learning,” *Syntax Literate : Jurnal Ilmiah Indonesia*, 6(1).
- Bambroo, P. dan Awasthi, A. (2021) “LegalDB: Long distilbert for legal document classification,” dalam *Proceedings of the 2021 1st International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies, ICAECT 2021*. Institute of Electrical and Electronics Engineers Inc. Tersedia pada: <https://doi.org/10.1109/ICAECT49130.2021.9392558>.
- Bintang, B. dkk. (2022) “Sustainable Development Goals (Sdgs): Kehidupan Sehat Dan Sejahtera Dalam Penanggulangan Covid-19 Di Daerah Semarang,” *Jurnal Pembangunan Berkelanjutan*, 5(1), hlm. 1–7. Tersedia pada: <https://doi.org/10.22437/jpb.v5i1.15563>.
- Devlin, J. dkk. (2019) “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm), hlm. 4171–4186.
- Dhina, M.M. dan Sumathi, S. (2022) “An innovative approach to classify hierarchical remarks with multi-class using BERT and customized naïve bayes classifier,” *International Journal of Engineering, Science and Technology*, 13(4), hlm. 32–45. Tersedia pada: <https://doi.org/10.4314/ijest.v13i4.4>.
- Elvy, E. dan Heriyanto, H. (2021) “Peran Perpustakaan Perguruan Tinggi Dalam Mendukung Implementasi Sustainable Development Goal 4,” *Baca J. Dokumentasi Dan Inf*, 42(1), hlm. 153.

- Erlin, E. *dkk.* (2022) “Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang,” *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 21(3), hlm. 677–690. Tersedia pada: <https://doi.org/10.30812/matrik.v21i3.1726>.
- Fajri, F. *dkk.* (2022) “Membandingkan Nilai Akurasi BERT dan DistilBERT pada Dataset Twitter Tahapan Penelitian,” *JUSIFO (Jurnal Sistem Informasi)*, 8(2), hlm. 71–80.
- Hanif, A.J. *dkk.* (2023) “Penerapan Natural Language Processing untuk Klasifikasi Bidang Minat berdasarkan Judul Tugas Akhir,” *Jurnal Sistim Informasi dan Teknologi*, 5(1), hlm. 41–49. Tersedia pada: <https://doi.org/10.37034/jsisfotek.v5i1.196>.
- Husin, N. (2023) “Komparasi Algoritma Random Forest, Naïve Bayes, dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN) Nanang Husin,” *Jurnal Esensi Infokom*, 7(1), hlm. 75.
- Irfan, M. (2021) *Named Entity Recognition Untuk Data Review Tempat Wisata Dengan Metode “Bidirectional Encoder Representations from Transformers.”*
- Jojoa, M. *dkk.* (2022) “Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization,” *AI & society*, hlm. 1–8.
- Joshi, A.P. dan Patel, B. V. (2021) “Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process,” *Oriental journal of computer science and technology* [Preprint]. Tersedia pada: <https://api.semanticscholar.org/CorpusID:241495446>.
- Julianda, A.R. dan Maharani, W. (2023) “Personality Detection on Reddit Using DistilBERT,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(5), hlm. 1140–1146. Tersedia pada: <https://doi.org/10.29207/resti.v7i5.5236>.
- Kurniawan, B. *dkk.* (2022) “Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (PSE) Menggunakan Algoritma Bidirectional Encoder Representations from Transformer (BERT),” *Jurnal Teknologi dan Sistem Informasi*, 3(4), hlm. 98–106.

- Mellyana, N. (2021) “Analisis penerapan konsep sustainable university dalam mendukung SDGs (studi kasus: pada dua universitas),” *Jurnal Pengelolaan Lingkungan Berkelanjutan (Journal of Environmental Sustainability Management)*, hlm. 799–815.
- Nurfatimah, S.A. dkk. (2022) “Membangun Kualitas Pendidikan di Indonesia dalam Mewujudkan Program Sustainable Development Goals (SDGs),” *Jurnal Basicedu*, 6(4), hlm. 6145–6154. Tersedia pada: <https://doi.org/10.31004/basicedu.v6i4.3183>.
- Pratiwi, V.R. dan Pardede, J. (2022) “Image Captioning Menggunakan Metode Inception-V3 dan Transformer,” *e-Proceeding FTI* [Preprint].
- Putri, M. dkk. (2023) “Studi Empiris Model BERT dan DistilBERT Analisis Sentimen pada Pemilihan Presiden Indonesia,” *Indonesian Journal of Computer Science*, 12(5).
- Saadah, F. dkk. (2023) “Klasifikasi Bidang Ilmu Pada Publikasi Terindeks GARUDA Kemdikbud Menggunakan Metode K-Nearest Neighbor (KNN),” *TRANSISTOR Elektro dan Informatika*, 5(2), hlm. 95–101.
- Sanh, V. dkk. (2019) “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108* [Preprint].
- Sulasminingsih, S. dkk. (2024) “Penerapan Tema SDGs Kehidupan Sehat dan Sejahtera untuk Menangani Polusi Udara di Jakarta,” *Jurnal Sains dan Teknologi*, 2024, 8.1: 18-26. [Preprint]. Tersedia pada: <https://doi.org/10.37817/ikraith-teknologi.v8i1>.
- Suryaningsum, S. (2020) “Strategi Universitas Meraih Nilai Tinggi Untuk Jurnal Terakreditasi Dalam Sinta,” *JSSH (Jurnal Sains Sosial dan Humaniora)*, 4(1), hlm. 73–79.
- Tandijaya, H.J. dkk. (2021) “Klasifikasi dalam Pembuatan Portal Berita Online dengan Menggunakan Metode BERT,” *Jurnal Infra*, 9(2), hlm. 320–325.
- Wijayanti, N.P.Y.T. dkk. (2021) “SMOTE: Potensi Dan Kekurangannya pada Survei,” *E-Jurnal Matematika*, 10(4), hlm. 235. Tersedia pada: <https://doi.org/10.24843/mtk.2021.v10.i04.p348>.