ANALISIS *WORD PREDICTION* DENGAN MENGGUNAKAN *LANGUAGE MODEL Bidirection Encoding Representations from Transformers* (BERT) PADA *DATASET* KALIMAT BAHASA INDONESIA

(Skripsi)

Oleh

AGHITA NAMIRA YULIZA
1917051019

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ANALISIS *WORD PREDICTION* DENGAN MENGGUNAKAN *LANGUAGE MODEL Bidirection Encoding Representations from Transformers* (BERT) PADA *DATASET* KALIMAT BAHASA INDONESIA

Oleh

AGHITA NAMIRA YULIZA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA KOMPUTER

Pada

Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRAK

ANALISIS *WORD PREDICTION* DENGAN MENGGUNAKAN *LANGUAGE MODEL Bidirection Encoding Representations from Transfordmers* (BERT) PADA *DATASET* KALIMAT BAHASA INDONESIA

Oleh

AGHITA NAMIRA YULIZA

Bahasa Indonesia sebagai bahasa nasional memiliki peran penting dalam berbagai bidang, termasuk pengembangan teknologi pemrosesan bahasa alami (*Natural Language Processing*). Salah satu pendekatan modern dalam NLP adalah penggunaan model transformer-based seperti BERT (*Bidirectional Encodder Representations from Transformers*) untuk menyelesaikan tugas *Masked Langugae Modeling* (MLM), yaitu menebak token yang hilang dalam suatu kalimat berdasarkan konteksnya. Tujuan penelitian ini adalah untuk mengevaluasi kinerja model BERT pada kalimat bahasa Indonesia dengan dataset 27.600 baris kalimat bahasa Indonesia. Model dilatih dengan dua skema, yaitu tanpa augmentasi (skema 1) dan dengan teknik augmentasi data (skema 2). Hasil evaluasi menunjukkan bahwa skema 2 memberikan kinerja yang lebih baik, dengan akurasi sebesar 42,1% (top-1), 53,7% (top-3), dan 58,1% (top-5), dibandingkan dengan skema 1 yang menghasilkan akurasi 29% (top-1), 42,6% (top-3), dan 52,6% (top-5). Peningkatan ini menunjukkan bahwa penggunaan augmentasi data dapat meningkatkan variasi kalimat dalam pelatihan model, kemampuan prediktif model terhadap kata-kata yang dimasking dapat ditingkatkan.

Kata kunci: *BERT, Masked Language Modelling, Bahasa Indonesia,* NLP*;*

ABSTRACT

WORD PREDICTION ANALYSIS USING THE BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) LANGUAGE MODEL ON INDONESIAN SENTENCE DATASETS

By

AGHITA NAMIRA YULIZA

Indonesian, as the national language, plays a crucial role in various fields, including the development of Natural Language Processing (NLP) technologies. One modern approach in NLP is the use of transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) to perform Masked Language Modeling (MLM), which involves predicting missing tokens in a sentence based on context. This study aims to evaluate the performance of the BERT model on Indonesian sentences using a dataset of 27,600 Indonesian sentence entries. The model was trained using two schemes: without augmentation (Scheme 1) and with data augmentation techniques (Scheme 2). Evaluation results show that Scheme 2 provides better performance, with an accuracy of 42.1% (top-1), 53.7% (top-3), and 58.1% (top-5), compared to Scheme 1 which achieved an accuracy of 29% (top-1), 42.6% (top-3), and 52.6% (top-5). This improvement indicates that data augmentation can enhance the diversity of training sentences, thereby improving the model's predictive capability for masked words.

***Keywords**: BERT, Masked Language Modeling, Indonesian Language, NLP*

Judul Skripsi	:	**ANALISIS *WORD PREDICTION* DENGAN MENGGUNAKAN *LANGUAGE MODEL Bidirection Encoding Representations from Transformers* (BERT) PADA *DATASET* KALIMAT BAHASA INDONESIA**
Nama Mahasiswa	:	Aghita Namira Yuliza
Nomor Pokok Mahasiswa	:	1917051019
Program Studi	:	S1 Ilmu Komputer
Jurusan	:	Ilmu Komputer
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Dr. rer. nat. Akmal Junaidi, M. Sc.
NIP. 19710129 199702 1 001

2. Ketua Jurusan Ilmu Komputer

Dwi Sakethi, S. Kom., M. Kom
NIP. 19680611 199802 1 001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. rer. nat. Akmal Junaidi, M.Sc.** .....

Sekretaris : **Favorisen R. Lumbanraja, Ph.D.** .....

Penguji Bukan Pembimbing : **Dewi Asiah Shofiana, S.Kom., M.Kom** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.

NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: **21 Mei 2025**

PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama : Aghita Namira Yuliza
NPM : 1917051019

Dengan ini menyatakan bahwa skripsi saya yang berjudul "**Analisis *Word Prediction* Dengan Menggunakan *Language Model Bidirection Encoding Representations From Transformers* (Bert) Pada *Dataset* Kalimat Bahasa Indonesia**" merupakan karya saya sendiri, bukan karya orang lain. Semua tulisan yang tertulis dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Jika di kemudian hari terbukti bahwa karya tulis ilmiah saya terbukti hasil menjiplak karya orang lain, maka saya siap menerima sanksi berupa pencabutan gelar yang saya peroleh.

Bandar Lampung, 5 Juni 2025

Aghita Namira Yuliza
NPM. 1917051019

RIWAYAT HIDUP

Penulis dilahirkan di Prabumulih pada tanggal 7 Agustus 2001, sebagai anak kedua dari dua bersaudara dari bapak Maiyucik dan Ibu Mareni Ferlin. Penulis menyelesaikan pendidikan formal di SD Negeri 07 Prabumulih dan selesai pada tahun 2013, dilanjutkan Pendidikan menengah pertama di SMPN 1 Prabumulih yang diselesaikan pada tahun 2016, dan melanjutkan ke Pendidikan menengah atas di SMAN 3 Prabumulih yang diselesaikan pada tahun 2019.

Pada tahun 2019 penulis terdaftar sebagai mahasiswa jurusan Ilmu Komputer Universitas Lampung melalui jalur SNMPT. Selama menjadi mahasiswa, penulis melakukan beberapa kegiatan antara lain.

1. Menjadi anggota bidang Sekretas Himpunan Mahasiswa Jurusan Ilmu Komputer pada periode 2019/2020
2. Menjadi Staff Ahli bidang Kementrian Dalam Negeri Badan Eksekutif Mahasiswa (BEM) Universitas Lampung pada periode 2020/2021
3. Menjadi Pembimbing Senior Divisi Kakak Asuh Program Orientasi Mahasiswa Jurusan Ilmu Komputer 2020
4. Menjadi Panita HPDD dan Pengelola akun media sosiak untuk acara PRJ Jurusan Ilmu Komputer 2020
5. Menjadi Panitia bidang Hubungan Masyarakat dalam seminar PT. Bukit Asam yang diselenggarakan BEM Universitas Lampung 2020
6. Mengikuti Program MBKM Jursan di Nusantara Regas Jakarta bidang Divisi Keuangan dan Sistem Informasi dari Juli – Desember 20222

7. Mengikuti KKN Periode 1 tahun 2022 di Desa Peninjauan, dan KWI di Lampung Timur tahun 2019
8. Asisten Dosen pada mata kuliah Matematika Diskrit dan Basis Data 2020/2021

MOTTO

“Sesungguhnya bersama kesulitan ada kemudahan.”
- (QS Al-Insyirah: 6) -

“Ketika Lelah, aku tidak berhenti. Aku istirahat dan Kembali berjuang”
- Author -

"Seperti BERT memahami dari dua arah, aku pun belajar dari kegagalan dan harapan."
- Author -

PERSEMBAHAN

Alhamdulilallahirobbilalamiin

Puji Syukur kepada Allah SWT atas berkah, Rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan penulisan skripsi ini. Sholawat serta salam tak lupa disanjungkan kepada nabi Muhammad SAW.

Kupersembahkan hasil karya ini kepada:

Ayah dan Mama

Terima kasih kuucapkan kepada Ayah dan Mama yang senantiasa penuh kesabaran, kasih sayang, dan cinta tulus yang selalu mengiringi perjalanan hidup penulis. Dan terima kasih banyak telah tanpa lelah mendidik dan membesarkan serta memberikan atas segala do'a dan motivasi yang tak pernah luput untuk menyemangati penulis dalam mencapai kesuksesan.

Ayuk dan Keluarga Besar

Terima kasih atas segala do'a -do'a baik, semangat, motivasi, kasih saying dan bantuan yang telah diberikan kepada penulis.

Sahabat dan Teman – Teman seperjuangan

Terima kasih telah memberikan semangat, dan motivasi kepada penulis selama perkuliahan. Terima kasih telah membantu penulis untuk belajar maupun berdiskusi dalam menyelesaikan perkuliahan hingga penyusunan karya yang sederhana ini.

Almamater Tercinta, Universitas Lampung dan Jurusan Ilmu Komputer

Tempat belajar mengemban seluruh ilmu untuk menjadi bekal hidup.

SANWACANA

Puji Syukur kepada Allah SWT yang telah memberikan Rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi dengan judul "Analisis *Word Prediction* dengan Menggunakan *Language Model Bidirection Encoding Representations from Transformers* (BERT) Pada *Dataset* Kalimat Bahasa Indonesia". Skripsi ini merupakan salah satu syarat dalam menyelesaikan proses perkulihan dan mendapatkan gelar Sarjana Komputer di Jurusan Ilmu Komputer FMIPA Universitas Lampung.

Selama proses penyusunan skripsi ini tidak terlepas dari dukungan dari banyak pihak, sehingga pada kesempatan ini penulis ingin mengucapkan rasa terima kasih kepada:

1. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku Dekan FMIPA Universitas Lampung.
2. Bapak Dwi Sakethi, M. Kom. selaku Ketua Jurusan Ilmu Komputer FMIPA Universitas Lampung.
3. Ibu Yunda Heningtyas, S. Kom,. M. Kom. selaku Sekretasi Jurusan Ilmu Komputer dan pembimbing akademik yang telah membimbing selama perkuliahan di Jurusan Ilmu Komputer.
4. Bapak Dr. rer. nat. Akmal Junaidi M. Sc. selaku pembimbing yang telah membimbingi saya dengan memberikan arahan, motivasi, kritik, dan saran dalam menyelesaikan skripsi ini dengan baik.
5. Bapak Favorisen R. Lumbanraja, Ph.D. selaku pembahas utama yang telah memberikan kritik dan saran yang sangat membantu dalam perbaikan skripsi ini.

6. Ibu Dewi Asiah Shofiana, S. Kom,. M. Kom. selaku pembahas pembantu yang telah memberikan kritik dan saran kepada penulis yang sangat membantu dalam perbaikan skripsi ini.
7. Kedua orang tua serta saudari kandung saya yang senantiasa memberikan doa, kepercayaan, dan dukungan baik moral atau materiel kepada penulis sehingga dapat menyelesaikan perkuliahan sampai skripsi dengan baik.
8. Ibu Ade Nora Maela, Bang Zainuddin, Mas Nofal, dan Mas Sam yang telah membantu segala urusan administrasi dan izin penulis di Jurusan Ilmu Komputer.
9. Bapak dan Ibu Dosen Jurusan Ilmu Komputer FMIPA Universitas Lampung yang telah memberikan ilmu bermanfaat.
10. Teman baik saya Indrias Meita Sari yang selalu memberikan semangat dan mendengarkan keluh kesah penulis selama proses perkuliahan sampai skripsi.
11. Teman-teman saya sejak awal perkuliahan Okta toyibah, Fanirizki Shofiana, Nur Ayu Octarina yang senantiasa memberikan semangat, dukungan, dan doa.
12. Teman baik saya Fakari yang selalu mendengarkan keluh kesah serta memberikan semangat, dukungan, dan doa kepada penulis.
13. Teman-teman saya di akhir perkuliahan Tasya Nursita Dewi, Revita Setia Ningsih, Melinda Sari yang senantiasa memberikan semangat dan doa kepada penulis.

Penulis berharap skripsi ini dapa menambah pengetahuan bagi pembaca tentang analisis kata prediksi bahasa Indonesia dengan metode BERT. Penulis meyadari bahwa dalam penulisan skripsi ini masih banyak kekurangan karena keterbatasan kemampuan, pengalaman, dan pengetahuan penulis. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan sebagai bahan evaluasi. Semoga skripsi ini dapat bermanfaat bagi semua pihak.

Bandar Lampung, 5 Juni 2025

Aghita Namira Yuliza

NPM. 1917051019

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR KODE PROGRAM

Kode Program Halaman

I. PENDAHULUAN

1.1 Latar Belakang

Negara Indonesia merupakan negara besar dengan jumlah penduduk 269 603,4 juta jiwa (Badan Pusat Statistik., 2020), terdiri dari negara kepulauan yang memiliki puluhan suku dengan bahasa yang berbeda – beda, Indonesia berhasil digiring untuk menerima satu bahasa di luar bahasa daerah mereka sebagai bahasa persatuan bangsa yaitu bahasa nasional. Bahasa Indonesia sebagai bahasa nasional yang berfungsi sebagai alat komunikasi mempunyai peran sebagai penyampai informasi (Zulfadhli *et al*., 2023). Penggunaan Bahasa Indonesia pada teknologi Bahasa alami atau *Natural Language Processing* (NLP) masih belum banyak digunakan (Putra *et al*., 2022). Analisis bahasa sangat penting dalam berbagai bidang seperti pendidikan, penelitian dan penerjemahan. Analisis bahasa Indonesia memiliki tingkat kesulitan seperti tata bahasa yang tidak teratur, variasi dialek dan keterbatasan data (Sa'diyah *et al*., 2023) . Tantangan – tantangan ini memerlukan pendekatan berbasis teknologi, khususnya dalam NLP untuk menghasilkan model yang lebih efektif dan akurat.

NLP merupakan salah satu bidang kecerdasan buatan dengan fungsi komunikasi antara manusia dan komputer melalui bahasa alami, teknologi NLP memungkinkan komputer dapat membaca, mendengar dan menafsirkan serta membuat keputusan (Rosyadi *et al*., 2020). Pemrosesan NLP merupakan bidang ilmu komputer yang berhubungan dengan kecerdasan buatan dan linguistik. Perkembangan NLP tentunya berhubungan dengan pendekatan *deep learning* dalam arsitekturnya. NLP memiliki fungsi memecah kata atau bahasa menjadi beberapa elemen yang lebih singkat, kemudian memahami hubungan antar elemen, serta mencari tahu

bagaimana elemen bisa bekerjasama dalam memasukan teks untuk berbagai tugas pemrosesan bahasa (Chandra *et al*., 2022.).

Prosedur *training* data dalam *deep learning* disebut juga *pre training*. Model *pre-training* dilakukan train pada *dataset* yang besar, salah satu metode *pre-training* yaitu BERT (*Bidirectional Encoder Representations from Transformers*). BERT merupakan *deep learning* model yang telah memberikan hasil canggih pada berbagai tugas NLP, termasuk analisis teks, klasifikasi, dan *question answering*. BERT memiliki 12 lapisan *Transformer* yang ditumpangkan di atas *encoder*, yang memungkinkan pemrosesan data yang lebih kompleks. Proses pelatihan BERT melibatkan pengaturan konfigurasi konfigurasi tinggi, waktu pelatihan yang banyak, pada tahun 2018 Devlin *et al*. mengusulkan sebuah model, yaitu BERT (*Bidirectional Encoder Representations from Transformers*) yang berhasil mendapatkan performa *state-of-the-art* pada banyak studi terkait NLP. BERT dirancang untuk mempelajari hubungan kontekstual anara kata – kata dalam sebuah teks secara dua arah (*bidirectional*), berbeda dari model sebelumnya yang hanya memahami konterks dari satu arah, yaitu kiri ke kanan atau kanan ke kiri. Pendekatan ini menggunakan teknik *Masked Language Model* (MLM). Dimana beberapa kata dalam teks disembunyikan secara acak, dan model dilatih untuk memprediksi kata-kata yang hilang tersebut berdasarkan konteks dari kedua arah (*bidirectional*). BERT menggunakan *Transformer* yang merupakan mekanisme yang mempelajari hubungan kontekstual antara kata – kata dalam teks menggunakan *self-attention mechanism.* Mekanisme ini memungkinkan model untuk memberikan perhatian pada hubungan setiap kata dalam kalimat, sehingga menghasilkan representasi yang kaya dan bermakna untuk setiap kata dalam konteksnya. (Putra *et al*., 2022).

Mask Language Model (MLM) merupakan salah satu *pre-trained* dari metode BERT dengan tujuan memprediksi kata yang ditutupi dengan konteks dari kedua arah (kiri dan kanan) dan mengevaluasi probabilitas urutan teks dari data (Wang *et al*., 2019). *Pre-trained* MLM BERT pertama kali dirilis dalam dua versi yaitu menggunakan korpus bahasa Inggris dan korpus Cina (Devlin *et al*., 2018.). Penelitian yang dilakukan oleh (Devlin *et al*., 2018) memanfaatkan MLM yang

didasarkan dengan mengubah urutan *input* dengan menghapus *token* dan kemudian melatih model untuk memprediksi token yang hilang tersebut, sehingga mampu merekonstruksi urutan aslinya. Strategi ini menjadikan BERT unggul dalam memahami hubungan semantik dan sintaksis dalam teks dibandingkan dengan pendekatan berbasis *unidirectional* yang digunakan oleh model – model sebelumnya, seperti GPT. Pengembangan model *pre-training* BERT dengan nama IndoBERT menggunakan korpus bahasa Indonesia untuk *pre-training* yang dilakukan oleh (Koto *et al*., 2020) telah berhasil melakukan *fine-tuning* untuk beberapa tugas pada pemrosesan bahasa alami seperti, *Next Tweet Prediction* dan *tweet ordering*.

Beberapa penelitian di atas, telah berhasil mengembangkan *language model* BERT pada beberapa bahasa. Akan tetapi, model yang dikembangkan menggunakan *dataset* dengan *size* yang besar sehingga memakan waktu dalam proses *training* dan juga harus mempunyai kapasitas *device* yang besar. Oleh karena itu, pada penelitian ini diajukan *language model* BERT untuk mengetahui hasil ketepatan prediksi kata dari *pre-trained* MLM menggunakan *dataset* kalimat bahasa Indonesia dengan kapasitas yang lebih kecil untuk mengurangi waktu dalam pemrosesan saat *training*.

1.2 Rumusan Masalah

Adapun rumusan masalah penelitian ini adalah sebagai berikut :

1. Bagaimana penerapan *language model* BERT pada *dataset* kalimat bahasa Indonesia untuk menentukan hasil *pre-trained* MLM?
2. Bagaimana tingkat akurasi *pre-trained* MLM untuk dataset kalimat bahasa Indonesia?

1.3 Batasan Masalah

Adapun Batasan Masalah dalam penelitian ini sebagai berikut :

1. Menggunakan *dataset* kalimat bahasa Indonesia yang didapatkan dari Kaggle.

2. Menggunakan *pre-trained* MLM pada *dataset* yang berjumlah 27600 baris kalimat.

1.4 Tujuan

Penelitian ini bertujuan untuk mengukur kinerja *pre-trained* MLM pada *dataset* kalimat bahasa Indonesia dengan *size* kapasitas yang lebih kecil.

1.5 Manfaat

Adapun manfaat penelitian ini adalah sebagai berikut :

1. Penelitian ini diharapkan dapat mengetahui hasil ketepatan prediksi kata yang di tutup atau disebut *Mask Language Model* dengan metode BERT menggunakan kalimat bahasa Indonesia yang *size dataset* lebih kecil.
2. Penelitian ini diharapkan dapat dijadikan referensi bagi peneliti lainnya untuk dikembangkan lebih lanjut dengan metode BERT atau lainnya menggunakan beberapa bahasa sebagai *dataset*-nya.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian yang dilakukan berlandaskan pada penelitian – penelitian terdahulu, sehingga memiliki kesamaan serta perbedaan pada objek penelitian. Tabel 1 menyajikan ringkasan dari penelitian terdahulu yang relevan dengan penelitian ini.

Tabel 1. Penelitian Terdahulu yang Terkait

No	Penelitian	Metode	Hasil
1	BERT *Pre-traning of Deep Bidirectional Transformers for Language Understanding* (Devlin *et al*, 2018)	*Bidirectional Encoder Representations from Transformers* (BERT).	Model berhasil mencapai hasil terbaru dari sebelas tugas NLP, termasuk meningkatkan: skor GLUE: 80,5% akurasi MultiNLI: 86,7% skor F1 pada tes SQuAD v1.1: 93,2 skor F1 pada tes SQuAD v2.0: 83,1
2	IndoLEM *and* IndoBERT*: A Benchmark Dataset and Pre-trained Language Model for Indonesian* NLP. (Koto *et al*, 2020)	IndoBERT	Akurasi *Next Tweet Prediction*: 93.7 *Tweet ordering* 0.59

Penelitian Tabel 1 mengacu pada penelitian – penelitian sebelumnya, yang dijelaskan berikut.

2.1.2 Bert *Pre-training of Deep Bidirectional Transformers for Language Understanding*

Penelitian ini dilakukan oleh (Devlin *et al*., 2018.) untuk mengetahui representasi bahasa terbaru yang disebut BERT, singkatan dari *Bidirectional Encoder Representations from Transformers*. Penelitian pertama kali dirilis dengan dua versi yaitu, korpus bahasa inggris dan korpus Cina. BERT dirancang untuk melakukan *pre-training* representasi dua arah mendalam dari teks yang tidak berlabel dengan mengkondisikan secara bersamaan pada konteks kiri dan kanan. Model ini mencapai hasil terbaik baru pada sebelas tugas pemrosesan bahasa alami, termasuk meningkatkan skor GLUE menjadi 80,5% (peningkatan absolut 7,7%), akurasi MultiNLI menjadi 86,7% (peningkatan absolut 4,6%), skor F1 pada tes SQuAD v1.1 menjadi 93,2 (peningkatan absolut 1,5 poin) dan skor F1 pada tes SQuAD v2.0 menjadi 83,1 (peningkatan absolut 5,1 poin).

2.1.3 IndoLem and IndoBERT: A *Benchmark Dataset and Pre-trained Language Model for Indonesian* NLP

Penelitian ini dilakukan oleh (Koto *et al*., 2020) untuk mengembangkan dan merilis IndoBERT, model bahasa BERT yang telah dilatih sebelumnya dalam satu bahasa Indonesia. Ini salah satu model BERT monolingual pertama untuk orang Indonesia bahasa, model ini mendapatkan hasil akurasi dari tugas *Next Tweet Prediction* bernilai 93,7 dan *tweet ordering* atau menyusun data dengan mengacak rangkaian pesan twitter dan menilai prediksi urutan dalam hal korelasi peringkat dengan aslinya yang didapatkan hasilnya sebesar 0,59.

2.2 Bahasa Indonesia

Bahasa merupakan sebuah simbol atau lambang bunyi yang berperan sebagai alat komunikasi. Bahasa digunakan masyarakat untuk bersosialisasi dengan satu sama

lain, sehingga peran bahasa sangat penting dalam kehidupan. Pada ruang lingkup yang kecil, masyarakat menggunakan bahasa daerah atau bahasa ibu, karena ada banyaknya daerah – daerah di Indonesia untuk mempersatu bangsa digunakan bahasa resmi atau bahasa nasional yaitu bahasa Indonesia. Bahasa Indonesia mempunyai kedudukan dan peranan penting bagi bangsa Indonesia dalam Negara Kesatuan Republik Indonesia. Bahasa Indonesia adalah media komunikasi utama bagi masyarakat Indonesia, penggunaan bahasa Indonesia diterapkan pada kehidupan sehari – hari masyarakat Indonesia (Febrianti *et al*., 2021).

2.3 Kalimat

Kalimat merupakan satuan bahasa terkecil yang terdiri dari beberapa kata yang memiliki makna. Kalimat adalah sarana komunikasi untuk menyampaikan pikiran atau gagasan kepada lawan bicara agar dapat dimengerti. Komunikasi akan berlangsung baik dengan kalimat yang disusun berdasarkan struktur yang benar, singkat, cermat, jelas maknanya, dan santun. Kalimat pada hakikatnya terdiri dari beberapa pola, penguasaan pola kalimat akan mempermudah pengguna bahasa dalam menyederhanakan kalimat. Kemudahan dalam menyampaikan ide dan memahami informasi yang dinyatakan oleh orang lain sehingga mengurangi kesalahpahaman dalam berkomunikasi (Aprilianto *et al*, 2017).

2.4 *Natural Language Processing* (NLP)

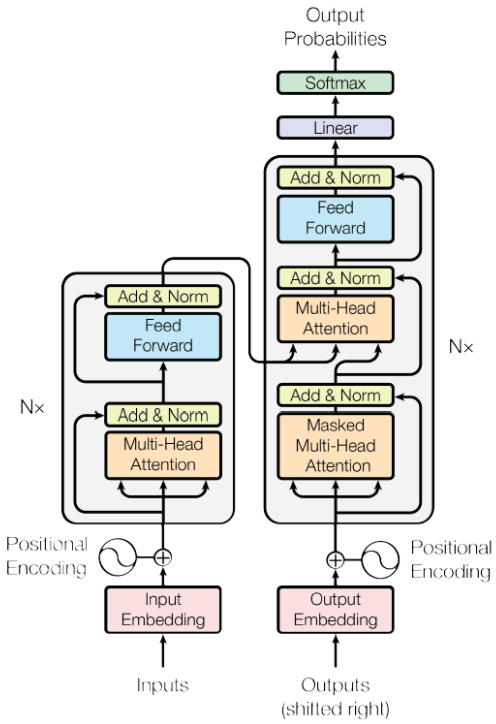
NLP merupakan bagian dari ilmu komputer dan ilmu linguistik yang mengkaji hubungan antara komputer dan bahasa alami (manusia). NLP merupakan cabang dari kecerdasan buatan (AI) yang berhubungan dengan melatih komputer agar dapat memahami, memproses, dan menghasilkan bahasa natural. Beberapa kajian NLP yaitu, segmentasi tuturan (*speech segmentation*), penandaan kelas kata (*part-of-speech tagging*), segmentasi teks (*text segmentation*), penentuan makna (*word sense disambiguation*). Secara umum NLP dibagi menjadi dua, yaitu *text-based application* dan *dialog-based* application. *Text-based application* merupakan aplikasi yang melakukan proses terhadap teks tertulis seperti dokumen, *e-mail*,

buku, dan lainnya. *Dialog-based application* merupakan bahasa lisan atau pengenalan suara, yang bisa juga memberikan interaksi dialog dengan mengetik teks pertanyaan melalui *keyboard* (Rohman *et al*., 2019).

NLP merupakan cabang ilmu AI yang berfokus pada pengolahan bahasa natural. Bahasa natural digunakan oleh manusia untuk komunikasi, bahasa alami membutuhkan waktu dalam memproses agar dapat dipahami dengan baik oleh komputer. Penerapan NLP pada beberapa aplikasi seperti *Chatbot* (digunakan untuk seolah – olah melakukan komunikasi antar pengguna dan komputer), *Lemmatization* dan *Stemming* (mengubah suatu kata pada bahasa tertentu pada bentuk dasar setiap kata suatu kalimat), *Text Summarization* (merupakan aplikasi peringkas teks dari bacaan), dan *Translation Tools* (aplikasi penerjemah bahasa) (Rumapea, 2021).

2.5 *Transformer*

Transformer merupakan sebuah arsitektur *deep learning* yang telah berhasil mengalami kemajuan besar di bidang pemrosesan bahasa alami (NLP) dan pengolahan data berstruktur. *Transformer* pertama kali diperkenalkan oleh Vaslin *et, al.* 2017. dalam makalah "*Attention Is All You Need*" yang diterbitkan oleh Konferensi *Neural Information Processing Systems* (NIPS) 2017. Arsitektur *Transformer* dikembangkan untuk memproses data teks yang didasarkan pada penggunaan mekanisme *attention* yang berupa urutan teks, sehingga model mengetahui hubungan antar dalam urutan. Model arsitektur *Transformer* terdiri dari blok *encoder* dan blok *decoder* yang masing – masing mempunyai *layer self-attention* dan *feed*-forward. Khusus pada *decoder* terdapat sebuah *layer attention* di antara dua *layer* tersebut yang berguna untuk membantu suatu *node* dalam mendapatkan konten inti yang membutuhkan attention. Ilustrasi model dari arsitektur *Transformer* dapat dilihat pada Gambar 1.



Gambar 1. Model Arsitektur *Transformer* (Vaswani *et al*., 2017)

2.6 *Bidirectional Encoder Representations from Transformers* (BERT)

BERT merupakan model *deep learning* dua arah yang berbasiskan arsitektur *Transformer*. Perbedaan BERT dan *deep learning* adalah penerapan pada *training Transformer* dua arah (*bidirectional*) ke dalam *language model*. BERT menggunakan teknik baru bernama *Masked Language Model* yang dapat melakukan *training dua arah*, dengan melihat urutan teks dari kiri ke kanan ataupun kombinasi dari kiri ke kanan dan kanan ke kiri. Transformer dapat belajar dan mengubah pemahaman yang diperoleh dari mekanisme *self-attention*. Pada dasarnya *Transformer* terdiri dari dua mekanisme, yaitu *encoder* dan *decoder*.

a. *Encoder*

Encoder digunakan untuk membaca data input eks. *Encoder* terdiri dari tumpukan N = 6 lapisan identik. Setiap lapisan memiliki dua sublapisan, lapisan *self-attention* dan jaringan saraf *feedforward*. Dengan lapisan *self-attention, encoder* dapat membantu node yang tidak hanya focus pada kata yang divisualisasikan, tetapi juga mendapatkan konteks dari kata tersebut.

Setiap posisi di *encoder* dapat memproses semua posisi lapisan sebelumnya di *encoder*.

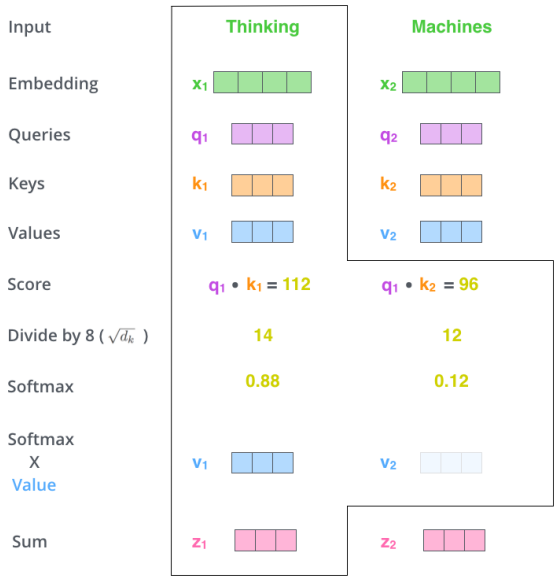
b. *Decoder*

Decoder bertugas menghasilkan urutan keluaran yang diprediksi. Komponen ini terdiri dari tumpukan N = 6 lapisan yang dapat diidentifikasi. Setiap lapisan memiliki dua sublapisan yang serupa dengan yang terdapat pada *encoder*. Ditambah *attention layer* di antara keduanya untuk membantu node mengambil informasi penting yang diinginkan dari *output encoder* melalui perhatian *multi-head*. Sema halnya dengan *encoder*, lapisan *self-attention* di *decoder* memungkinkan setiap posisi di *decoder* untuk menangani semua posisi sebelumnya dan posisi saat ini dalam urutan.

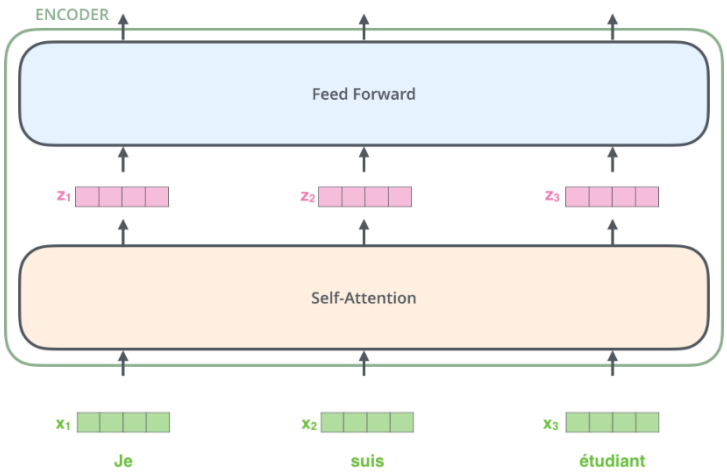
Langkah – Langkah berikut adalah proses yang berjalan pada *encoder* dan *decoder* pada gambar 1 (Khan *et al*., 2022) :

1. Setiap istilah kata yang masuk ke dalam *encoder* diubah menjadi vector menggukann teknik *embedding.* Lantaran *self-attention layer* tidak memiliki kemampuan untuk mengenali urutan istilah – istilah dalam sebuah kalimat. *Positional encoding* ditambahkan untuk memberikan informasi posisi berdasarkan istilah – istilahnya. Setiap vektor menurut *input* mempunyai ukuran dimensi 512. Proses ini hanya diterapkan pada *encoder* yang berada paling bawah (pertama), sebagai akibatnya *encoder* lainnya akan mendapat hasil menurut *encoder* yang pertama.
2. *Input vector* kemudian melewati dua *layer* utama dalam setiap *encoder* yaitu *self-attention layer* dan *feed-forward neural network*. Pada tahap *self-attention layer* dibentuk tiga vektor menurut masing – masing input vector yaitu *Query, Key* dan *Value vector*. Ketiga vektor ini dibentuk menggunakan mengalikan *embedding*. Masing-masing vector memiliki dimensi 64. Nilai *self-attention* untuk setiap kata dihitung menggunakan mengalikan *query vector* dan *key vector* misalnya yang terdapat dalam Gambar 2. Kemudian, nilai *self-attention* dibagi 8, karena 8 merupakan akar kuadrat menurut dimensi tiap vector yaitu 64. Nilai *self-attention* juga dihitung menggunakan softmax. Sebagai akibatnya tiap *value vector* akan dikali menggunakan nilai

menurut softmax. Akhirnya *value vector* dijumlahkan dan sebagai hasil menurut *self-attention layer*. *Output* menurut *self-attention layer* lalu masuk ke *feed-forwarrd* untuk masing – masing posisi misalnya yang tertera dalam Gambar 3.



Gambar 2. Proses pada *self-attention* (Alammar, 2018)



Gambar 3. Proses pada *Encoder* (Alammar, 2018)

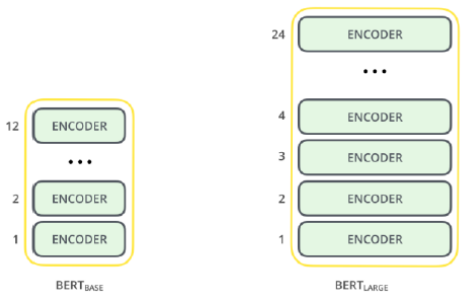
3. Setelah seluruh proses pada *encoder* selesai, *output* dari *encoder* yaitu *vector key* dan *vector value*, masuk ke dalam *decoder*. Setiap input dan *output* dari *encoder* dan *decoder self-awareness layer* dan *feed-forward neural network* akan melalui tahapan *add* dan *norm*, yang mencakup struktur residual dan

lapisan normalisasi. Proses yang dilakukan oleh *decoder* serupa dengan *encoder*, namun memiliki lapisan *self-attention* tambahan yang berada di antara *layer self-aware* dan jaringan saraf *feed-forward* yang membantu *decoder* untuk fokus pada bagian kata yang relevan. *Layer self-aware decoder* hanya bisa peduli dengan posisi sebelum *output. Output* dari setiap langkah terus disuplai ke *decoder* dan *output decoder* sama dengan output *encoder*. Akhirnya, *output* dari tumpukan *decoder* menghasilkan vektor dengan nilai *floating point*. Untuk memasukkannya ke dalam kata-kata, dibutuhkan lapisan tambahan dari lapisan yang terhubung penuh, bersama dengan lapisan *softmax*.

Model BERT memiliki arsitektur berupa *multi-layer bidirectional transformer* mirip seperti yang dilakukan pada implementasi asli transformer, namun hanya memanfaatkan bagian *encoder* saja. BERT tersedia dalam dua varian ukuran, yaitu BERT-*base* dan BERT-*large*.

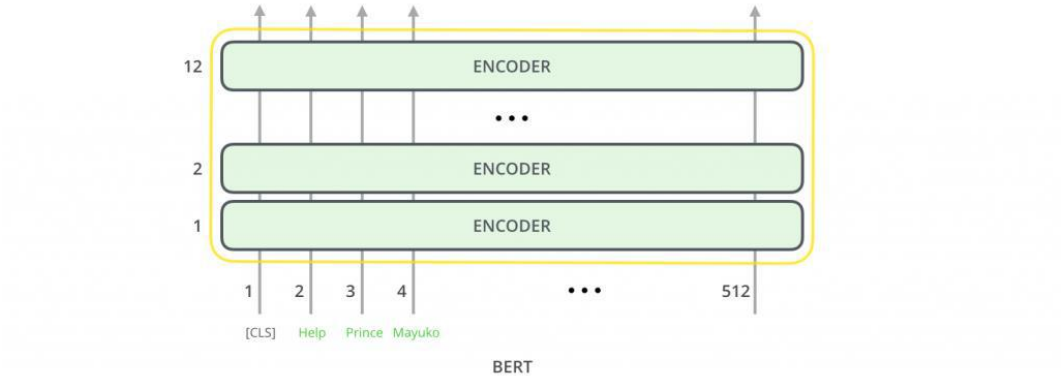
Google awalnya merilis dua versi ukuran model, yaitu BERTBASE dan BERTLARGE. L mempresentasikan jumlah dari lapisan *Transformer*, H mempresentasikan *hidden size*, dan A mempresentasikan jumlah dari *multi-head attention*. Perbedaan antara BERTBASE dan BERTLARGE dapat dilihat pada Gambar 4.

1. BERTBASE: L = 12, H = 768, A = 12, Total Parameter = 110M.
2. BERTLARGE: L = 24, H = 1024, A = 16, Total Parameter = 340 M.



Gambar 4. Perbedaan BERTBASE dan BERTLARGE (Alammar, 2018).

Pada implementasinya, BERT tersedia dalam dua versi model, yaitu BERTBASE dan BERTLARGE seperti yang ditunjukkan pada Gambar 4. Kedua ukuran versi model BERT ini memiliki jumlah lapisan *encoder* atau *Transformer Blocks* yang berbeda. BERTBASE memiliki *encoder* dengan 12 *layers*, 12 *self-attentions* heads, hidden size sebesar 768, dan 110M parameters. BERTLARGE terdapat 24 *layers encoders*, 16 self-*attention heads*, *hidden size* sebesar 1024, dan 340M parameter. BERTBASE dilatih selama 4 hari menggunakan 4 *cloud* TPUs sedangkan BERTLARGE membutuhkan 4 hari pelatihan menggunakan 16 TPUs.



Gambar 5. Arsitektur BERT (Alammar, 2018).

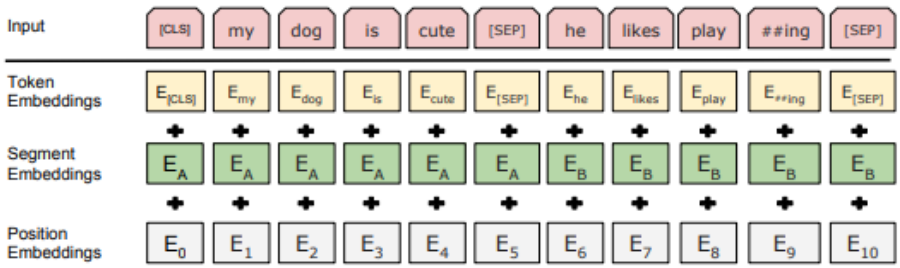
Sesuai dengan namanya, BERT hanya menggunakan *encoder*, sehingga arsitektur BERT terlihat seperti Gambar 5. BERT berbeda dengan model terarah (*directional*) yang melihat urutan teks dari kiri-ke-kanan, kanan-ke-kiri, atau gabungan dari kiri-ke-kanan dan kanan-ke-kiri. Model bahasa yang dilatih secara *bidirectional* dapat memiliki pemahaman yang lebih dalam tentang konteks daripada model bahasa satu arah.

Representasi *input* BERT ditampilkan pada Gambar 6. Berikut merupakan langkah-langkah tokenisasi dalam BERT (Khalid, 2019):

1. Tokenisasi: Membagi teks menjadi unit-unit kecil yang disebut token, biasanya berupa kata atau bagian dari kata. BERT menggunakan metode tokenisasi WordPiece, dimana satu token dapat dibagi lagi menjadi sub-token.
2. *Token Embeddings*: BERT menambahkan dua token khusus, yaitu [CLS] di awal kalimat dan [SEP] di akhir. Token [CLS] berfungsi sebagai representasi

keseluruhan kalimat, sementara [SEP] digunakan untuk memisahkan kalimat saat terdapat lebih dari satu kalimat dalam satu *input*.

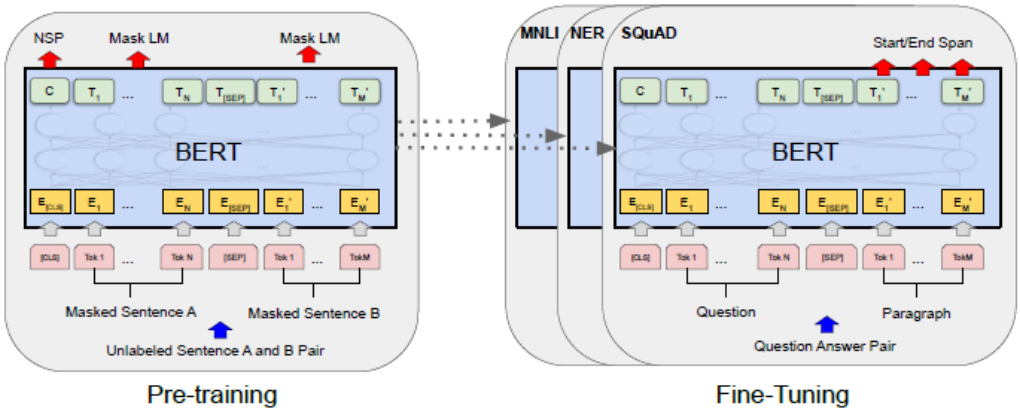
3. Konversi Token menjadi ID: Setiap token dalam *input* kemudian dikonversi menjadi ID numerik yang sesuai berdasarkan kamus token yang telah ditentukan. Selanjutnya, setiap ID token dikonversi menjadi vektor dengan mengambil nilai *embedding* dari matriks *embedding* kata yang telah dilatih sebelumnya. Matriks *embedding* menggambarkan setiap kata dalam ruang vektor yang terdiri dari banyak dimensi.
4. *Segment Embeddings*: Bila *input* terdiri dari dua kalimat, masing-masing token harus diidentifikasi berasal dari kalimat pertama atau kedua. Hal ini dilakukan dengan memberikan ID segmen, yaitu 0 untuk kalimat pertama dan 1 untuk kalimat kedua.
5. *Position Embedding*: Untuk menambahkan informasi posisi dalam urutan token, BERT menggabungkan embedding posisi ke dalam setiap token. Embedding ini berupa vektor posisi yang sudah ditentukan sebelumnya dan ditambahkan ke token embeddings guna menyampaikan urutan kata dalam kalimat.



Gambar 6. Representasi Input BERT (Devlin *et al*., 2018)

Model BERT terdapat dua fase penggunaan yaitu *pre-training* dan *fine-tuning* yang ditunjukan pada Gambar 7. *Pre-training* termasuk *unsupervised learning* karena model dilatih menggunakan data yang tidak berlabel untuk mengekstrak pola. sedangkan *fine-tuning* data diinisialisasi dengan parameter *pre-trained* dan datanya menggunakan data yang berlabel untuk mengoptimalkan kinerja model pada tugas tersebut. *Fine-tuning* dilakukan dengan menambahkan lapisan khusus untuk tugas di atas model BERT yang sudah dilatih sebelumnya dan kemudian melatih seluruh model dari awal hingga akhir pada data khusus tugas tersebut. Jumlah parameter di

lapisan khusus tugas jauh lebih kecil dari model BERT yang sudah dilatih sebelumnya. Selama *fine-tuning*, model dilatih dengan tingkat pembelajaran yang lebih kecil dibandingkan saat *pre-training*. Hal ini karena model yang sudah dilatih sebelumnya telah belajar fitur umum bahasa dan lapisan khusus tugas perlu mempelajari hanya fitur khusus dari tugas *downstream* (Sun *et al*., 2019). Fase *pre-trained* dari BERT terdiri dari dua *unsupervised tasks*, yaitu *Masked language Model* dan *Next Sentence Prediction*).



Gambar 7. BERT *pre-training* dan *fine-tuning* (Devlin *et al.*, 2018)

a. *Masked language Model* (MLM)

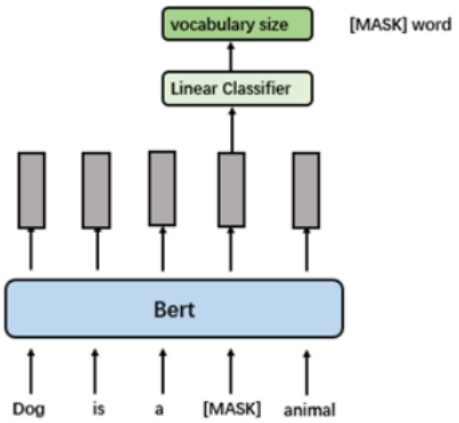
   Tugas MLM adalah memberi *mask* atau menutupi 15% kata – kata ke setiap urutan token *input* secara acak dan kemudian diprediksi nilai asli dari token yang ditutupi.

b. *Next Sentence Prediction* (NSP)

   Dalam proses *pre-trained*, BERT dapat menerima pasangan kalimat sebagai *input* dan dilatih untuk memprediksi jika pasangan kalimat tersebut merupakan kalimat berikutnya atau hanya satu kalimat saja. Jika terdapat kalimat A dan B, maka 50% BERT menganggap bahwa kalimat B adalah kalimat berikutnya dari kalimat A. 50% lainnya adalah kalimat acak yang diambil dari korpus.

BERT menggunakan fungsi dua arah dan mekanisme *multi-layer attention* Pada pelaksanaan *train* sebuah representasi *bidirectional*, token yang di *input* akan di-

masked secara acak pada sebagian persentase token tersebut dan token yang telah di-*masked* akan diprediksi. Tujuan dari penambahan *pre-training* adalah banyaknya tugas NLP seperti QA yang perlu memahami relasi antar dua kalimat (Devlin *et al*., 2018.)



Gambar 8. BERT *pre-training flowchart* (Hu *et al*., 2021)

Gambar 8 merupakan proses *Masked Language Modeling*, dimana model memasukkan kalimat bertopeng ke dalam BERT, dan mengklasifikasikan vektor posisi yang sesuai dengan masker dalam matriks keluaran model, dan labelnya adalah subskrip yang sesuai dengan kata bertopeng dalam kamus (Hu *et al*., 2021). MLM menjadi teknik dasar dalam bidang pemrosesan bahasa alami, khususnya sejak diperkenalkan model BERT (*Potential of masked*). BERT memilih rasio *masking* 15% berdasarkan alasan bahwa model tidak dapat mempelajari representasi yang baik ketika terlalu banyak teks yang tersamar, dan pelatihan tidak efisien ketika terlalu sedikit yang tersamar. Di sisi lain, MLM mengalami biaya komputasi yang signifikan karena hanya mempelajari 15% token per urutan, sedangkan Language Model *autoregresif* memprediksi setiap token dalam urutan (Wettig *et al*., 2022).

2.7 *Hyperparameter*

Hyperparameter adalah variabel yang mempengaruhi *output* model. Nilai dari *hyperparameter* ini tidak bisa diubah selama proses model dioptimasi, yang berarti nilai tidak tergantung pada data (Gunawan *et al*, 2020). *Hyperparameter* yang dilakukan saat proses *pre-trained* BERT meliputi beberapa hal yaitu *batch size*, *learning rate*, jumlah *epoch*, dan sebagainya (Devlin *et al*, 2018).

2.8.1 *Batch Size*

Batch size merupakan jumlah sampel yang diolah di setiap iterasi saat *pre-trained*. *Batch size* yang terlalu kecil dapat menghasilkan waktu pelatihan lebih lama, sedangkan berukuran besar akan memengaruhi kebutuhan memori dan juga semakin membutuhkan waktu pelatihan *dataset* (Palakodati *et al*, 2020).

2.7.2 *Learning Rate*

Learning Rate adalah parameter yang memantau seberapa besar perubahan yang dilakukan pada bobot model selama pelatihan. Nilai *learning rate* yang terlalu tinggi dapat menyebabkan pelatihan menjadi tidak memusat, sedangkan nilai *learning rate* yang terlalu rendah bisa menyebabkan pelatihan menjadi lambat dan butuh waktu yang lama (Rismiyati dan Luthfiarta, 2021).

2.7.3 Jumlah *Epoch*

Jumlah *epoch* menentukan jumlah iterasi yang akan dilakukan pada seluruh data pelatihan. Jumlah *epoch* yang sedikit akan mengakibatkan model tidak terlatih dengan baik, sedangkan jumlah yang terlalu banyak bisa menyebabkan *overfitting*. *Epoch* merupakan suatu kondisi di mana setiap *dataset* yang dimasukkan ke model telah selesai melewati seluruh proses pelatihan pada *neural network* dalam satu kali putaran secara *forward propagation* dan *back propagation* (Rozaqi dan Sunyoto, 2020).

2.8 *Top-k*

Metode top-k dipilih setelah dihitung nilai bobot dari logit yang dikonversi menjadi probabilitas dengan perhitungan *softmax* (Wibowo & Indriyawati, 2020). Fungsi *softmax* digunakan untuk mengubah vektor logit menjadi fungsi probabilitas, fungsi menghasilkan nilai antara 0 dan 1. Pengaturan MLM untuk menghasilkan logit bagi token yang diperlukan untuk membangun kembali semua entitas (Brayne A *et al*, 2022). Persamaan mengubah logit menjadi probabilitas dengan perhitungan *softmax*:

$$Softmax\ (z^{(k)}) = \frac{e^{z^{(k)}}}{\sum_{i=1}^{K} e^{z^{(i)}}} \text{ untuk k = 1, …, V} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Keterangan :

$z^{(k)}$ adalah logit atau skor mentah dari model untuk token top-k

V adalah jumlah total token dalam *vocabulary* (32.000 token)

$e^{z^{(k)}}$adalah ekponensial dari logit $z^{(k)}$, digunakan untuk memastikan nilai positif.

$\sum_{i=1}^{K} e^{z^{(i)}}$adalah penjumlahan semua eksponensial logits untuk normalisasi

2.9 *Top-k Accuracy*

Top-k *accuracy* adalah metrik evaluasi yang mengukur seberapa baik model dalam memprediksi label yang benar berdasarlam kecocokannya yang direpresntasikan oleh nilai k sebagai batas peringkat. Jika label yang benar termasuk dalam k prediksi tersebut, maka prediksi dianggap benar. Metrik ini memberikan gambaran lebih baik tentang kinerja model, yang menghasilkan beberapa kemungkinan untuk setiap input. Top-k *accuracy* tidak hanya melihat ketepatan hasil, tetapi juga relevansi rekomendasi prediksi label yang sering muncul dalam rekomendasi k teratas (Dhimas *et al*., 2025).

Hasil dari logit yang diubah menjadi probabilitas ini diterapkan pada perhitungan top-k, menangkap label yang benar di antara k dengan probabilitas tertinggi (Giovannotti *et al*., 2021). Persamaan menunjukkan perhitungan *accuracy* prediksi:

$$Accuracy = \frac{Prediksi\ Benar}{Total\ Data\ uji\ Mask} \times 100\ \% \quad \text{.....(2)}$$

III. METODELOGI PENELITIAN

3.1 Waktu dan Tempat

Penelitian dilakukan di Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA), Jalan Prof. Dr. Ir. Sumantri Brojonegoro No.1 Gedong Meneng, Kecamatan Rajabasa, Kota Bandar Lampung, pada Maret 2023 sampai dengan Februari 2025. Waktu penelitian yang dilakukan akan dijelaskan pada Tabel 2.

Tabel 2. Waktu Penelitian

Nama Kegiatan	Maret 2023	April 2023–Desember 2023	Januari 2024 – Oktober 2024	November 2024 – Februari 2025
Persiapan Dataset				
Preprocessing Dataset				
Model BERT				
Pengujian Model BERT				
Analisis Hasil				

3.2 Alat

Alat yang digunakan dalam penelitian ini adalah sebagai berikut:

3.2.1 Perangkat Keras (*Hardware)*

Perangkat keras yang digunakan dalam penelitian ini adalah menggunakan laptop ASUS dengan spesifikasi RAM 8.00 GB, 512 GB SSD, *processor* Intel Core i7-

8565U dengan GPU Intel UHD Graphics 620 dan komputer NVIDIA GTX 1080 Ti.

3.2.2 Perangkat Lunak (*Software*)

Perangkat lunak yang digunakan dalam penelitian ini diantaranya menggunakan sistem operasi Ubuntu dan Jupyter Notebook sebagai lembar kerja untuk mengeksekusi kode program.

3.3 *Dataset*

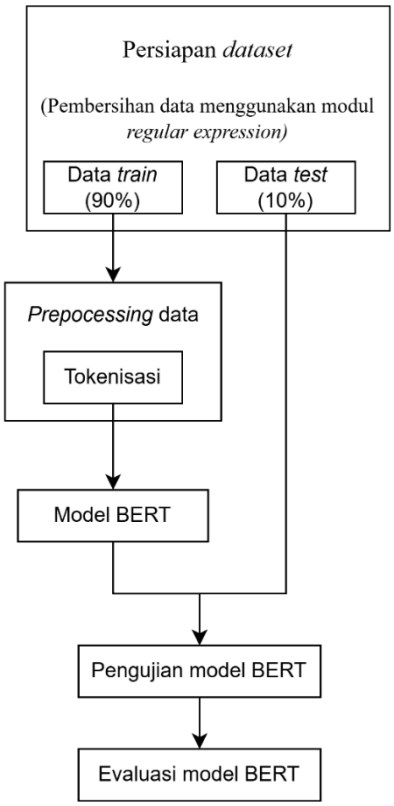
Penelitian ini dilakukan menggunakan *dataset* kalimat bahasa Indonesia. Kalimat bahasa Indonesia yang diperoleh dari Kaggle berupa kalimat – kalimat pendek, kalimat pendek tersebut terdiri dari beberapa kata. *Dataset* kalimat bahasa Indonesia memiliki 27600 baris. Berikut merupakan contoh *dataset*:

Tabel 3. Contoh *dataset*

ayah biasanya membantu tetangga dengan pekerjaan rumah
saya merasa seperti boneka yang telah disalahgunakan
aku tidak akan berhenti
saya membeli sepasang sepatu baru
saya merasa seperti saya akan berjuang dan gagal dan menderita dan menjadi benar benar bodoh
kamu biasanya membawa bekal makan siang dari rumah ke sekolah
saya menjadi lebih kesal ketika bruce sedikit lebih lelah dari pekerjaan dari biasanya saya merasa sedikit ditolak
saya merasa seperti orang bodoh yang mengatakan bahwa karena saudara perempuan saya yang manis telah melalui tahun terburuk dalam hidupnya pada saat yang sama
saya merasa lebih kacau daripada yang pernah saya alami dan itu tidak ada hubungannya dengan pekerjaan sekolah saya
saya merasa sedikit stress

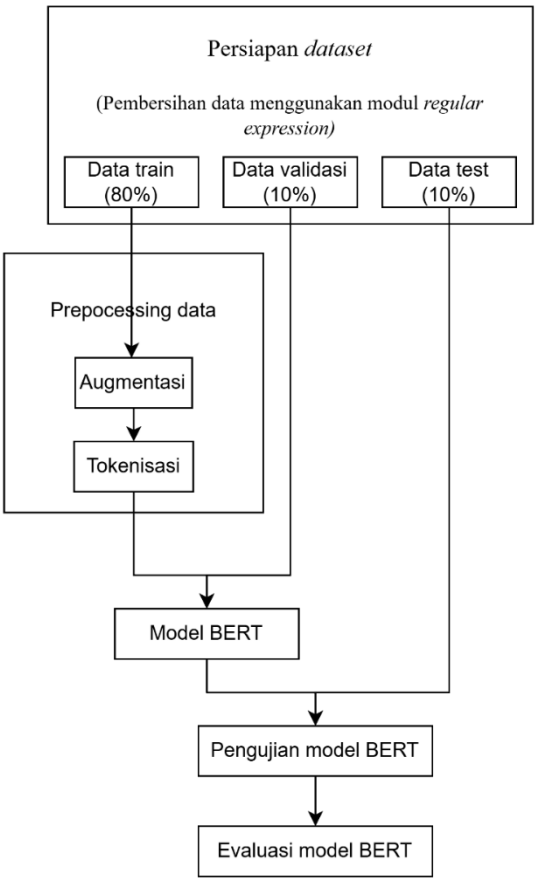
3.4 Metode Penelitian

Adapun alur metode penelitian dalam pengerjaan penelitian ini terdiri dari dua skema penelitian, skema pertama dapat dilihat pada Gambar 9.



Gambar 9. Alur metode penelitian skema 1

Gambar 9 menjelaskan alur metode penelitian skema pertama. *Dataset* yang sudah dibersihkan dengan modul *regular expression* langsung dibagi menjadi data *training* dan data *testing*. Proporsi data *training* dipilih secara *random* sebanyak 90%, sisa 10% data digunakan untuk data *testing*.



Gambar 10. Alur metode penelitian skema 2

Skema kedua dapat dilihat pada Gambar 10. Perbedaan mendasar skema ini dengan skema pertama adalah *dataset*, selain untuk *training* dan *testing*, juga dialokasikan untuk validasi sebesar 10%. Dengan alokasi yang baru ini, distribusi *dataset* menjadi 80%, 10%, dan 10% masing-masing untuk data *training*, data validasi, dan data *testing*.

3.4.1 Persiapan *Dataset*

Dataset sebelum digunakan dalam pemrosesan *training*, *dataset* tersebut terlebih dahulu melalui tahap pembersihan (*data cleaning*) untuk memastikan kualitas dan konsistensinya. Proses pembersihan meliputi penghapusan karakter yang tidak perlu, seperti simbol atau tanda baca, atau memperbaiki susunan baris yang belum tepat, serta memperbaiki setiap awal kalimat dengan huruf kecil.

Tahap pembersihan selesai, *dataset* dibagi sesuai dengan skema penelitian yang dirancang. Pada skema 1, *dataset* dibagi menjadi dua bagian utama, yaitu data pelatihan (*training set*) sebesar 90% dan data pengujian (*testing set*) sebesar 10%. Skema 2, *dataset* dibagi menjadi tiga bagian utama, yaitu data pelatihan (*training set*) sebesar 80%, data validasi sebesar 10%, dan data pengujian sebesar 10%. Pembagian dataset ini dilakukan untuk memastikan bahwa model dapat diuji pada data yang belum pernah dilihat sebelumnya, sehingga hasil evaluasi dapat mencerminkan kemampuan generalisasi model terhadap data baru.

3.4.2 *Preprocessing* Data

Pada tahap ini dilakukan *preprocessing* data untuk menyiapkan *dataset* sebelum digunakan dalam proses pelatihan model. Penelitian dengan skema satu tidak mengalami augmentasi, sedangkan skema dua mengalami augmentasi pada data pelatihan. Augmentasi data yang digunakan menggunakan metode *random swap* dan *random deletion*, setiap kalimat asli ditambahkan dua kalimat augmentasi. Augmentasi bertujuan meningkatkan keragaman variasi struktur kalimat, dengan mengganti urutan kata ataupun menghapus kata dari kalimat asli tersebut. Dalam *preprocessing data* ada tokenisasi teks, yaitu proses memecah setiap kalimat menjadi token-token yang dapat diproses oleh model. Tokenisasi dilakukan menggunakan *tokenizer* BERT, sebuah fungsi yang dirancang untuk memecah teks sesuai dengan *vocabulary* BERT. Tokenisasi yang dilakukan pada penelitian skema 1 menggunakan data pelatihan yang tidak mengalami augmentasi data, sedangkan skema 2 menggunakan data pelatihan yang telah melalui proses augmentasi data.

Selama proses tokenisasi, *tokenizer* BERT juga menambahkan token khusus, seperti [CLS] di awal setiap kalimat untuk mempresentasikan konteks keseluruhan, dan [SEP] untuk menandai akhir dari setiap kalimat atau segmen.

3.4.3 Model BERT

Penelitian ini menggunakan model BERT sebagai pendekatan utama untuk mempelajari kontekstual antar kata dalam kalimat bahasa Indonesia, *dataset* dilatih dengan salah satu tugas dari *pre-trained* BERT yaitu *Mask Language Model* (MLM). Model BERT menyediakan beberapa *hyperparameter* untuk mengoptimalkan kinerja model yang dilakukan. *hyperparameter* yang digunakan dalam penelitian yaitu *batch size, epoch, learning rate, number head attention, vocab size, intermediate size, hidden size*.

3.4.4 Pengujian Model BERT

Penelitian dengan kalimat bahasa Indonesia setelah melalui proses *precocessing* dan *training,* selanjutnya penelitian melakukan pengujian terhadap model BERT yang telah dilatih menggunakan tugas *Masked Language Model* (MLM). Pengujian dilakukan menggunakan *dataset* kalimat bahasa Indonesia yang sebelumnya telah diproses dan disiapkan secara khusus untuk menguji kemampuan prediksi model.

Pengujian model yang telah melalui proses *pre-trained* MLM dilakukan dengan cara menutupi salah satu kata secara acak dari baris kalimat. Kata yang di*mask* ini digantikan oleh token khusus [MASK], yang kemudian diproses oleh model BERT. Model menganalisis kalimat yang telah dimodifikasi dengan mempertimbangkan konteks dari semua kata lain yang ada dalam kalimat, baik sebelum atau setelah kata yang di*mask*. Berdasarkan informasi kontekstual, model mencoba memprediksi token asli dari kata yang telah di*mask*. Proses ini bertujuan untuk mengevaluasi performa model dalam memahami dan menggeneralisasi pola dan keterkaitan antar kata dalam suatu kalimat.

3.4.5 Analisis Hasil

Langkah terakhir dalam penelitian yang dilakukan adalah melakukan analisis hasil yang bertujuan untuk mengetahui kekurangan dan kelebihan model yang digunakan. Hasil analisis dapat digunakan untuk meningkatkan kualitas model.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Penelitian ini memiliki hasil penelitian yang dapat disimpulkan sebagai berikut.

1. Penerapan *language model* BERT pada *dataset* kalimat bahasa Indonesia dilakukan dengan melatih model sendiri menggunakan *Masked Language Model* (MLM) sebagai metode *pre-training*. Proses pelatihan skema 1 mencakup pengaturan konfigurasi 20 *epoch*, 32 *batch size*, 0,00001 *optimizer*, sementara skema 2 dengan teknik augmentasi dan pengaturan konfigurasi 10 *epoch*, 64 *batch size*, dan 0,00001 *optimizer*. *Tokenizer* yang digunakan adalah cahya/bert-base-indonesian-1.5G. Model yang dilatih dilakukan pengujian dengan kalimat mengandung [MASK], kemudian model memprediksi kata yang paling sesuai.
2. Hasil pengujian model menunjukkan bahwa skema 2 jauh lebih baik daripada skema 1, akurasi top-1 yang awal hanya mencapai 29% dengan menggunakan teknik augmentasi akurasi meningkat menjadi 42,1%, yang berarti model dapat menebak kata lebih baik. Namun, ketika evaluasi diperluas ke top-3, akurasi skema 1 yang bernilai 42,6% menjadi meningkat ketika menggunakan model skema 2 dengan nilai 53,7%, ini menunjukkan bahwa model mampu menghasilkan kandidat kata yang cukup relevan meskipun tidak selalu berada di posisi pertama. Selanjutnya, pada top-5 akurasi skema 1 meningkat hingga 52,6% dan skema 2 jauh lebih meningkat menjadi 58,1%, ini menyatakan model memiliki pemahaman konteks yang lebih baik ketika diberikan lebih banyak pilihan kata dan dengan augmentasi data membuat model lebih banyak variasi pada kalimatnya. Peningkatan akurasi ini menunjukkan bahwa meskipun model belum optimal dalam prediksi satu kata paling tepat, model

tetap memiliki potensi dalam memahami struktur bahasa Indonesia dan menghasilkan kata – kata yang relevan dalam prediksi top-k.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan, diperoleh beberapa saran untuk menjadi bahan pertimbangan penelitian masa mendatang diantaranya:

1. *Dataset* yang digunakan untuk data pelatihan dapat diperbanyak agar model memiliki pemahaman yang lebih luas terhadap variasi bahasa dan struktur kalimat dalam bahasa Indonesia
2. Melakukan *fine-tuning* pada *domain* spesifik untuk meningkatkan akurasi prediksi kata sesuai konteks tertentu
3. Melakukan pelabelan data atau *post tagging* agar model dapat berlatih dengan baik dan menghasilkan hasil yang lebih baik

DAFTAR PUSTAKA

Alammar, J. (2018). The Illustrated Transformer. Https://Jalammar.Github.Io/Illustrated-Transformer/

Aprilianto, T., & Badawi, A. (2017). Sistem Koreksi Kata Dan Pengenalan Struktur Kalimat Berbahasa Indonesia Dengan Pendekatan Kamus Berbasis Levenshtein Distance. *Jurnal SPIRIT*, 9(1), 48-61.

Badan Pusat Statistik. (2020). Jumlah Penduduk Hasil Proyeksi Menurut Provinsi dan Jenis Kelamin (Ribu Jiwa), 2018 – 2020. https://www.bps.go.id/indicator/12/1886/1/jumlah-penduduk-hasil-proyeksi-menurut-provinsi-dan-jenis-kelamin.html

Brayne, A., & Wiatrak, M. (2022.). On Masked Language Models for Contextual Link Prediction. *Proceedings of Deep Learning Inside Out*, 87-99.

Chandra, A. A., Nathaniel, V., Satura, F. R., Dharma Adhinata, F., & Studi, P. (2022). Pengembangan Chatbot Informasi Mahasiswa Berbasis Telegram Dengan Metode Natural Language Processing. *Jurnal ICTEE*, 3(1), 20–27.

Dhimas, D., Putra, P., Swanjaya, D., & Ramdhani, R. A. (2025). Sistem Rekomendasi Laptop Menggunakan Metode Collaborative Filltering Dan Weighted Product Pada Toko Online Indojaya Computer, *Prosiding Seminar Nasional Teknologi Dan Sains*, 4.

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). Bert: Pre-Training Of Deep Bidirectional Transformers For Language Understanding. Https://Github.Com/Tensorflow/Tensor2tensor

Eka Rosyadi, H., Amrullah, F., David Marcus, R., & Rahman Affandi, R. (2020). 619 Rancang Bangun Chatbot Informasi Lowongan Pekerjaan Berbasis Whatsapp Dengan Metode NLP (Natural Language Processing). BRILIANT: *Jurnal Riset Dan Konseptual*, 5.

Febrianti, Y. F., Pulungan, R., Bahasa, P., Indonesia, S., & Al-Washliyah, U. M. (2021). Penggunaan Bahasa Gaul Terhadap Eksistensi Bahasa Indonesia Pada Masyarakat. *Jurnal Ilmu Pendidikan*, 2(1).

Gunawan, M. I., Sugiarto, D., & Mardianto, I. (2020). Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Search pada

Algoritma Logistic Regression. *Jurnal Edukasi dan Penelitian Informatika*, 6(3).

Giovannotti, P., Holloway, R., Gammerman, A., Carlsson, L., Luo, Z., Cherubin, G., & Nguyen, K. A. (2021). Transformer-Based Conformal Predictors For Paraphrase Detection. *Proceedings Of Machine Learning Research*, 152.

Harahap, R. N., Muslim, K., & Korespondensi, P. (2020). Peningkatan Akurasi Pada Prediksi Kepribadian Mbti Pengguna Twitter Menggunakan Augmentasi Data. *Jurnal Teknologi Informasi dan Ilmu Komputer* (JTIIK), 7(4), 815-822. https://doi.org/10.25126/jtiik.202073622

Hu, P., Dong, L., & Zhan, Y. (2021). Bert Pre-Training Acceleration Algorithm Based On Mask Mechanism. *Journal Of Physics: Conference Series*, 2025(1). Https://Doi.Org/10.1088/1742-6596/2025/1/012038

Iskandar Zulkarnain Maulana Putra, T., Farhan Bukhori, A., Ilmu Pengetahuan Alam, Dan, & Gadjah Mada, U. (2022). Model Klasifikasi Berbasis Multiclass Classification Dengan Kombinasi Indobert Embedding Dan Long Short-Term Memory Untuk Tweet Berbahasa Indonesia (Classification Model Based On Multiclass Classification With A Combination Of Indobert Embedding And Long Short-Term Memory For Indonesian-Language Tweets). *Jurnal Ilmu Siber Dan Teknologi Digital* (JISTED), 1(1), 1–28. Https://Doi.Org/10.35912/Jisted.V1i1.1509

Khalid, S. (2019). Bert Explained: A Complete Guide With Theory And Tutorial. Towards Machine Learning. Https://Towardsml.Wordpress.Com/2019/09/17/Bert-Explained-A-Complete-Guide-With-Theory-And-Tutorial/

Khan, M., Naeem, M. R., Al-Ammar, E. A., Ko, W., Vettikalladi, H., & Ahmad, I. (2022). Power Forecasting Of Regional Wind Farms Via Variational Auto-Encoder And Deep Hybrid Transfer Learning. Electronics (Switzerland). 11(2). Https://Doi.Org/10.3390/Electronics11020206

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). Indolem And Indobert: A Benchmark Dataset And Pre-Trained Language Model For Indonesian NLP. Http://Arxiv.Org/Abs/2011.00677

Mas, R., Panca, R. W., Atmaja1, K., & Yustanti2, W. (2021). Analisis Sentimen Customer Review Aplikasi Ruang Guru Dengan Metode Bert (Bidirectional Encoder Representations From Transformers). *JEISBI*, 02.

Nofiyanti, E., & Oki Nur Haryanto, E. M. (2021). Analisis Sentimen Terhadap Penanggulangan Bencana Di Indonesia. *Jurnal Ilmiah Sinus*, 19(2), 17. Https://Doi.Org/10.30646/Sinus.V19i2.563

Palakodati, S,S,S, *et al,* (2020) 'Fresh and Rotten Fruits Classification Using CNN and Transfer Learning'. *Revue d'Intelligence Artificielle*, 34(5), 617–622. DOI: 10,18280/ria,340512,

Putra, T. I. Z. M., Suprapto., Arif, F. B. 2022. Model Klasifikasi Berbasis *Multiclass Classification* dengan Kombinasi Indobert *Embedding* dan *Long Short-Term Memory* untuk *Tweet* Berbahasa Indonesia. *Jurnal Ilmu Siber Dan Teknologi Digital (JISTED)*. 1(1): 1 – 28.

Rismiyati, R, and Luthfiarta, A, (2021) 'VGG16 Transfer Learning Architecture for Salak Fruit Quality Classification', *Telematika: Jurnal Informatika Dan Teknologi Informasi*, 18(1), p, 37, DOI: 10,31315/telematika,v18i1,4025,

Rohman, A. N., Utami, E., & Raharjo, S. (2019). Deteksi Kondisi Emosi Pada Media Sosial Menggunakan Pendekatan Leksikon Dan Natural Language Processing. *Jurnal Eksplora Informatika*, 9(1), 70–76. Https://Doi.Org/10.30864/Eksplora.V9i1.277

Rozaqi, A,J, and Sunyoto, A, (2020) 'Identification of Disease in Potato Leaves Using Convolutional Neural Network (CNN) Algorithm', In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, IEEE, pp, 72–76, DOI: 10,1109/ICOIACT50329,2020,9332037,

Rumapea, H. (2021). Akuntansi Deteksi Kemiripan Artikel Melalui Keywords Dengan Metode Fuzzy String Matching Dalam Natural Language Processing. *Jurnal METHOMIKA: Manajemen Informatika Komputerisasi* 5(1). Https://Doi.Org/10.46880/Jmika.Vol5no1.Pp60-66

Sa'diyah, L. S., & Sri, A. S. 2023. Alternatif Strategi Pembelajaran Melalui Kegiatan Bercerita Dan Dramatisasi Kreatif Dalam Pembelajaran Berbahasa Lisan. *Jurnal Bahasa Sastra dan Pengajrannya*. 7(1), 86-96.

Situmorang, R., Rahayu, W. I., Nuraini, R., & Fathonah, S. (2023). Model Algoritma K-Nearest Neighbor (K-Nn) Dan Naïve Bayes Untuk Prediksi Kelulusan Mahasiswa. *Jurnal Mahasiswa Teknik Informatika*, 7(1).

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How To Fine-Tune Bert For Text Classification? Http://Arxiv.Org/Abs/1905.05583

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need.

Wang, A., & Cho, K. (2019). Bert Has A Mouth, And It Must Speak: Bert As A Markov Random Field Language Model. Http://Arxiv.Org/Abs/1902.04094

Wettig, A., Gao, T., Zhong, Z., & Chen, D. (2022). Should You Mask 15% In Masked Language Modeling? Http://Arxiv.Org/Abs/2202.08005

Wibowo, R., & Indriyawati, H. (2020). Top-k Feature Selection untuk Deteksi Penyakit Hepatitis Menggunakan Algortime Naïve Bayes 1 Fakultas

Teknologi Informasi dan Komunikasi. *Universitas Semarang, Indonesia Jl. Soekarno Hatta Tlogosari*, 11(1), 1-9.

Zulfadhli, M., Anshori, D. S., & Sunendar, D. (2023). Kebijakan Pembelajaran Mkwk Bahasa Indonesia Di Perguruan Tinggi: Implementasi Dan Tantangannya. *SEMANTIK*, 12(1), 125-130.