ELANG: Journal of Interdisciplinary Research
E-ISSN: 3025-2482

BERT dan Bahasa Indonesia: Studi tentang Efektivitas Model NLP Berbasis *Transformer*

Mukhlis Amien[1]*, Go Frendi Gunawan[2]
[1]amien@stiki.ac.id, [2]gofrendiasgard@gmail.com

[1,2] *Informatika, Sekolah Tinggi Informatika & Komputer Indonesia, Malang, Indonesia*

Informasi Artikel
Diterima: 25-01-2024
Direvisi: 30-01-2024
Diterbitkan: 30-01-2024

Kata Kunci
BERT; NLP; NLP Bahasa Indonesia; NER; POS Tagging; Model Transformer

****) Author Korespondensi***
amien@stiki.ac.id

Abstrak
Penelitian ini menyelidiki efektivitas model BERT dalam pemrosesan bahasa Indonesia, sebuah bahasa dengan struktur linguistik yang unik. Masalah penelitian dari studi ini berfokus pada evaluasi adaptasi model NLP berbasis transformer, khususnya BERT, untuk bahasa Indonesia dan membandingkan performanya dengan aplikasi pada bahasa lain. Metodologi penelitian ini yaitu mengungkapkan potensi model BERT dalam mengatasi tantangan linguistik khas bahasa Indonesia, memberikan wawasan penting tentang peningkatan performa model dalam konteks bahasa yang beragam. Hasil Penelitian kami menunjukkan bahwa metode BERT untuk bahasa Indonesia mendapatkan akurasi 96.8% untuk operasi POS Tagging, akurasi 74,9% untuk operasi NER UGM, dan 90,1% untuk operasi NER UI .

Abstract
This research investigates the effectiveness of the BERT model in translating Indonesian, a language with a unique linguistic structure. The research problem of this study focuses on evaluating the adaptation of transformer-based NLP models, especially BERT, for Indonesian and comparing its performance with applications in other languages. This research methodology reveals the potential of the BERT model in overcoming linguistic challenges unique to Indonesian, providing important insights into improving model performance in diverse language contexts. Our research results show that the BERT method for Indonesian gets 96.8% accuracy for POS Tagging operations, 74.9% accuracy for UGM NER operations, and 90.1% for UI NER operations.

1. Pendahuluan

Bahasa Indonesia, yang digunakan oleh lebih dari 275,7 juta orang, memainkan peran krusial dalam komunikasi dan teknologi NLP (Natural Language Processing). Uniknya, Bahasa Indonesia memiliki kekayaan linguistik dan tantangan spesifik, seperti variasi dialek dan struktur bahasa yang kompleks, yang belum sepenuhnya dieksplorasi dalam studi NLP. Ini menimbulkan kebutuhan untuk mempelajari penerapan model NLP berbasis transformer seperti BERT (Bidirectional Encoder Representations from Transformers), terutama dalam meningkatkan pemrosesan dan pemahaman bahasa ini. Studi sebelumnya telah menunjukkan kemajuan signifikan dalam NLP menggunakan BERT pada berbagai bahasa, namun masih terdapat ruang untuk eksplorasi lebih dalam terutama untuk Bahasa Indonesia (Sebastian, Purnomo, & Sembiring, 2022).

Salah satu tantangan utama dalam NLP untuk Bahasa Indonesia adalah keterbatasan dataset yang berkualitas dan representatif. Hal ini diperparah oleh kurangnya studi yang menyeluruh tentang bagaimana model transformer seperti BERT dapat diadaptasi dan dioptimalkan untuk Bahasa Indonesia. Penelitian ini akan fokus pada identifikasi dan penanganan masalah tersebut, dan bagaimana solusi ini dapat dibandingkan dengan penggunaan BERT pada bahasa lain, seperti Belanda dan Cina (de Vries et al., 2019); (Cui et al., 2019).

Penelitian ini bertujuan untuk secara mendalam memahami dan meningkatkan efektivitas BERT dalam konteks Bahasa Indonesia, dengan fokus khusus pada mengatasi tantangan yang dihadapi dalam NLP Bahasa Indonesia. Penelitian ini bertujuan untuk tidak hanya mengembangkan model BERT yang lebih efektif untuk Bahasa Indonesia, tetapi juga untuk menghasilkan wawasan yang dapat diterapkan dalam peningkatan model NLP berbasis transformer untuk bahasa lain, memberikan kontribusi penting pada penelitian global di bidang ini (Perez & Reinauer, 2022); (Zhang, Ding, Yu, O'Uchi, & Fujita, 2022).

Dengan menyelaraskan pendekatan penelitian ini dengan tantangan unik yang dihadapi oleh Bahasa Indonesia dalam bidang NLP, diharapkan dapat memberikan kontribusi signifikan pada pemahaman dan pengembangan teknologi NLP yang lebih efektif dan inklusif.

Pengembangan NLP untuk Bahasa Indonesia telah menarik perhatian peneliti dengan penerapan model-model seperti BERT yang telah sukses diterapkan pada bahasa-bahasa dengan sumber daya besar. Dalam konteks Bahasa Indonesia, penelitian oleh Rini Wijayanti mengeksplorasi transfer learning berbasis cross-lingual word embedding untuk peringkasan teks ekstraktif, menyoroti potensi cross-lingual transfer learning sebagai solusi bagi bahasa dengan sumber daya rendah seperti Bahasa Indonesia (Wijayanti, 2023).

Selanjutnya, ulasan oleh Audrey pada platform Cash AI menguraikan secara mendalam mengenai mekanisme kerja BERT, termasuk konsep token, segment, dan positional embeddings yang memungkinkan pemahaman yang lebih kaya terhadap konteks kalimat. Penjelasan ini memberikan dasar yang kuat bagi peneliti yang ingin menerapkan BERT pada bahasa-bahasa dengan tantangan unik, termasuk Bahasa Indonesia (Audrey, 2023).

Analisis dampak praktis dari penerapan BERT dalam NLP, sebagaimana dijelaskan dalam sumber yang sama, menyediakan wawasan tentang potensi aplikasi teknologi ini dalam pemrosesan bahasa alami, dari pengenalan entitas nama hingga analisis sentimen dan pengembangan sistem Q&A, yang semuanya relevan dengan pemrosesan Bahasa Indonesia (Audrey, 2023).

Penelitian ini akan membangun atas dasar penelitian-penelitian terdahulu dengan fokus pada adaptasi dan optimisasi BERT untuk Bahasa Indonesia, bertujuan untuk mengatasi beberapa keterbatasan yang dihadapi dalam studi-studi sebelumnya, seperti ketersediaan dataset berkualitas dan representatif untuk bahasa dengan sumber daya terbatas.

Dengan memahami dan mengevaluasi kemajuan yang telah dibuat serta tantangan yang masih ada, penelitian ini berupaya tidak hanya untuk memperkaya literatur yang ada mengenai penggunaan BERT untuk Bahasa Indonesia tetapi juga untuk memajukan praktik NLP secara keseluruhan dengan menyesuaikan teknologi ini untuk memenuhi kebutuhan spesifik dari Bahasa Indonesia.

NLP telah berkembang pesat dalam dekade terakhir, terutama karena pengenalan model pra-latih seperti BERT, yang telah merevolusi cara pemrosesan dan pemahaman bahasa dalam bidang ini. Seperti yang dijelaskan oleh Paass dan Giesselbach (2023), NLP kini berada pada fase di mana model-model foundation, termasuk BERT, memberikan dasar yang kuat untuk berbagai aplikasi NLP, mulai dari pemahaman kontekstual hingga generasi teks. Sementara BERT sendiri telah menjadi model transformatif dalam NLP, penelitian terkini, seperti yang dibahas oleh Nozza, Bianchi, dan Hovy (2020), telah menunjukkan bagaimana adaptasi BERT ke bahasa-bahasa tertentu dapat membantu mengatasi tantangan linguistik yang unik. Misalnya, BERT yang disesuaikan untuk bahasa Persia, seperti ParsBERT yang dijelaskan oleh Farahani et al. (2020), menunjukkan keefektifan BERT dalam konteks linguistik yang berbeda. Hal ini membuktikan fleksibilitas dan adaptabilitas model BERT, tidak hanya dalam mengatasi tantangan linguistik, tetapi juga

dalam aplikasi yang lebih luas, seperti yang ditunjukkan dalam penelitian oleh Perez dan Reinauer (2022), di mana BERT diintegrasikan ke dalam klasifikasi teks dengan menggunakan analisis data topologi. Kesimpulannya, BERT tidak hanya telah merevolusi NLP, tetapi juga terus memberikan inovasi dan solusi untuk berbagai tantangan dalam pemrosesan bahasa, seperti yang diuraikan dalam survei oleh Min et al. (2021), menyajikan perjalanan terbaru dalam pemahaman bahasa dengan model-model bahasa besar dalam NLP.

Bahasa Indonesia menawarkan tantangan unik dalam ranah NLP, tidak hanya karena kekayaan variasi dialek dan struktur bahasa yang kompleks tetapi juga karena kurangnya representasi dalam penelitian NLP global. Variasi dialek, sebagaimana dijelaskan oleh Sebastian, Purnomo, dan Sembiring (2022), menciptakan kesulitan dalam pemodelan NLP karena perbedaan signifikan dalam kosakata dan tata bahasa antar wilayah. Lebih lanjut, struktur bahasa yang melibatkan penggunaan partikel dan afiksasi menambah kerumitan dalam menginterpretasikan konteks semantik dan sintaksis dalam teks Bahasa Indonesia. Contohnya, dalam aplikasi seperti analisis sentimen, variasi dialek dapat menyebabkan interpretasi yang berbeda terhadap emosi atau nuansa yang sama, sehingga memerlukan pendekatan yang lebih spesifik dan adaptif.

Solusi dan teknologi terkini untuk mengatasi tantangan ini termasuk pengembangan model NLP yang lebih berfokus pada karakteristik linguistik unik Bahasa Indonesia. Hal ini mencakup pemanfaatan teknik pembelajaran mesin yang canggih dan adaptasi model bahasa besar seperti BERT yang telah dilatih pada berbagai bahasa untuk memperoleh pemahaman kontekstual yang lebih mendalam. Penting juga untuk memperhatikan aspek sosiolinguistik Bahasa Indonesia, seperti penggunaan bahasa formal dan informal, yang memiliki implikasi besar dalam pemahaman konteks dan makna.

Pendekatan multibahasa dan multi dialek dalam NLP memungkinkan pemahaman yang lebih baik tentang konteks linguistik yang lebih luas dan membantu dalam mengatasi kesenjangan penelitian yang muncul akibat fokus yang berlebihan pada bahasa-bahasa dominan seperti Inggris atau Mandarin. Melalui pendekatan ini, teknologi NLP dapat menjadi lebih inklusif dan akurat dalam memproses dan memahami keunikan Bahasa Indonesia.

Penerapan BERT dalam berbagai bahasa dan konteks telah menunjukkan hasil yang signifikan dalam peningkatan performa tugas-tugas pemrosesan bahasa alami (NLP). Dalam konteks bahasa Belanda, model RoBERTa yang berbasis BERT, seperti RobBERT, telah menunjukkan peningkatan yang signifikan dalam berbagai tugas NLP untuk Bahasa Belanda, terutama dalam menangani dataset yang lebih kecil (Delobelle, Winters, & Berendt, 2020). Penelitian lain oleh de Vries et al. (2019) mengembangkan dan mengevaluasi model BERT monolingual untuk Bahasa Belanda, BERTje, yang mengungguli model BERT multibahasa pada berbagai tugas NLP. Model ini didasarkan pada dataset besar dan beragam, yang menunjukkan pentingnya data pelatihan yang luas dan bervariasi dalam pengembangan model BERT untuk bahasa tertentu.

Untuk Bahasa Cina, BERT juga telah diadaptasi dan diaplikasikan dengan sukses. Cui et al. (2019) memperkenalkan strategi whole word masking (wwm) untuk BERT Bahasa Cina, menghasilkan model MacBERT yang meningkatkan performa pada berbagai tugas NLP Cina. Penelitian ini menunjukkan bahwa strategi pelatihan dan adaptasi khusus untuk Bahasa Cina dapat memberikan peningkatan yang signifikan dalam performa model.

Dalam upaya mengadaptasi teknologi pemrosesan bahasa alami (NLP) untuk Bahasa Indonesia, model BERT (Bidirectional Encoder Representations from Transformers) telah memainkan peran penting. Sebastian, Purnomo, dan Sembiring (2022) melakukan penelitian komprehensif mengenai penerapan BERT dalam Bahasa Indonesia, meliputi analisis sentimen, klasifikasi, dan ringkasan teks. Mereka menyoroti keunikan linguistik Bahasa Indonesia, seperti morfologi yang kompleks dan struktur sintaksis, yang memengaruhi implementasi model BERT. Penelitian ini penting dalam memahami bagaimana model BERT beradaptasi dengan kekhasan linguistik Bahasa Indonesia, yang berbeda dari bahasa lainnya (Sebastian, Purnomo, & Sembiring, 2022).

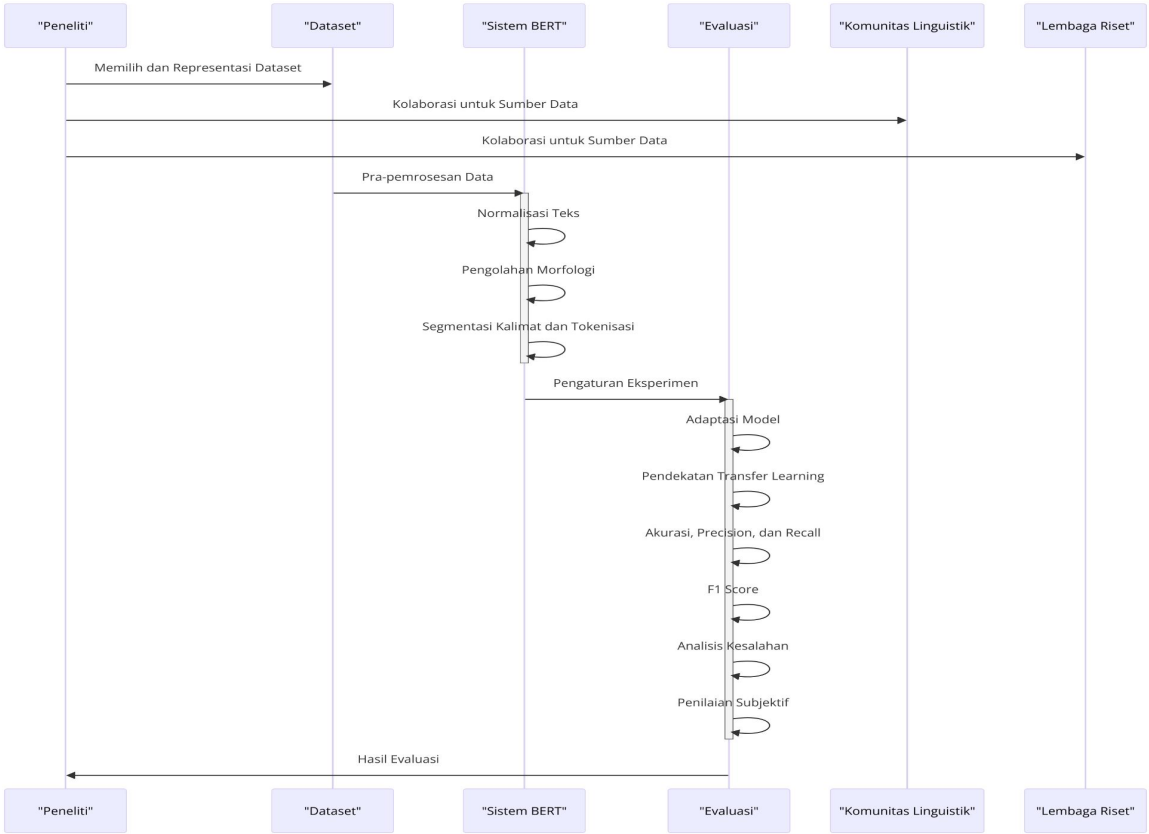
Selanjutnya, Suwarningsih et al. (2022) mengkaji kinerja model BERT dan variasinya, seperti ALBERT dan ELECTRA, dalam konteks Bahasa Indonesia, khususnya untuk pengembangan sistem tanya jawab. Penelitian ini berkontribusi pada pemahaman tentang bagaimana dataset berkualitas tinggi dan representatif dapat memperkuat pengembangan model BERT untuk Bahasa Indonesia. Mereka menyajikan contoh konkret strategi adaptasi dan optimalisasi yang berhasil, memberikan wawasan yang lebih mendalam tentang pendekatan yang efektif dalam konteks ini (Suwarningsih et al., 2022).

Koto, Lau, dan Baldwin (2021) memperkenalkan IndoBERTweet, model BERT pra-latih untuk Twitter Indonesia, yang memberikan perspektif baru dalam penggunaan BERT untuk konteks Bahasa Indonesia. Studi ini mengeksplorasi bagaimana adaptasi dan inisialisasi kosakata yang efektif dapat meningkatkan kinerja BERT dalam konteks Bahasa Indonesia, memberikan kontribusi penting pada pemahaman tentang bagaimana model BERT dapat dioptimalkan untuk kebutuhan bahasa tertentu (Koto, Lau, & Baldwin, 2021).

Nityasya et al. (2022) melakukan penelitian tentang distilasi pengetahuan dari model BERT-base ke berbagai model student. Mereka memberikan analisis kritis mengenai bagaimana dataset dan metode pelatihan memengaruhi kinerja model dalam bahasa Indonesia. Penelitian ini memberikan pemahaman yang lebih mendalam tentang tantangan dalam mengadaptasi BERT untuk Bahasa Indonesia, serta strategi pelatihan yang efektif (Nityasya et al., 2022).

Terakhir, Le et al. (2022) mengeksplorasi penerapan model multibahasa yang berbasis pada BERT dan ELECTRA untuk pemahaman bahasa alami, termasuk Bahasa Indonesia. Studi ini memberikan wawasan tentang implikasi jangka panjang tantangan ini terhadap pengembangan NLP di Indonesia, termasuk dampaknya terhadap teknologi yang berorientasi pada pengguna Bahasa Indonesia (Le et al., 2022).

2. Metodologi Penelitian



***Gambar 1**. Diagram alur metodologi penelitian*

Diagram alur yang disajikan menunjukkan proses metodologi penelitian yang rinci, melibatkan beberapa entitas dan langkah-langkah yang berinteraksi selama siklus penelitian. Dalam diagram ini, "Peneliti" memulai proses dengan memilih dan merepresentasikan dataset yang akan digunakan. Peneliti juga berkolaborasi dengan "Komunitas Linguistik" dan "Lembaga Riset" untuk mendapatkan akses ke sumber data yang berkualitas. Kemudian, "Dataset" tersebut masuk ke fase pra-pemrosesan yang dilakukan oleh "Sistem BERT". Di sini, sistem melaksanakan langkah-langkah seperti normalisasi teks, pengolahan morfologi, serta segmentasi kalimat dan tokenisasi. Setelah pra-pemrosesan, sistem mengatur eksperimen yang meliputi penyesuaian model dan penerapan teknik transfer learning. Fase "Evaluasi" kemudian mengambil alih dengan mengevaluasi model menggunakan berbagai metrik, termasuk akurasi, presisi, recall, skor F1, analisis kesalahan, dan penilaian subjektif. Akhirnya, hasil evaluasi tersebut dikembalikan ke "Peneliti".

2.1. Data dan Sumber Data

Untuk memahami dan meningkatkan efektivitas BERT dalam konteks Bahasa Indonesia, penelitian ini akan menggunakan dataset yang berkualitas dan representatif. Ini melibatkan:

1. Pemilihan Dataset: Mengumpulkan dataset yang kaya dan beragam dari sumber-sumber seperti media sosial, teks berita, dan literatur, untuk menangkap variasi linguistik Bahasa Indonesia. Fokus pada dataset yang mencakup dialek dan register bahasa yang berbeda akan membantu dalam memahami kompleksitas linguistik Bahasa Indonesia.
2. Representasi Dataset: Memastikan bahwa dataset mencakup spektrum sosiolinguistik yang luas dari Bahasa Indonesia, termasuk penggunaan bahasa formal dan informal serta variasi dialek.

2.2. Proses Pra-pemrosesan Data

Proses pra-pemrosesan data akan disesuaikan untuk mengatasi tantangan spesifik Bahasa Indonesia, termasuk:

1. Normalisasi Teks: Membersihkan dataset dari noise dan mengonversi teks ke format standar, termasuk menghilangkan slang dan mengganti singkatan dengan bentuk penuhnya.
2. Pengolahan Morfologi: Menggunakan teknik-teknik pemrosesan morfologis untuk menangani afiksasi dan pembentukan kata dalam Bahasa Indonesia.
3. Segmentasi Kalimat dan Tokenisasi: Mengembangkan atau menggunakan alat tokenisasi yang efektif untuk Bahasa Indonesia, mempertimbangkan struktur kalimat dan partikel khas bahasa.

2.3. Pengaturan Eksperimen

Pengaturan eksperimental akan dirancang untuk mengatasi keterbatasan teknologi dan sumber daya, serta untuk mengevaluasi efektivitas BERT untuk Bahasa Indonesia:

1. Pengaturan Hardware dan Software: Memanfaatkan sumber daya komputasi yang tersedia secara optimal, termasuk cloud computing dan GPU, untuk pelatihan dan evaluasi model.
2. Adaptasi Model: Menyesuaikan pengaturan model BERT (seperti jumlah lapisan, ukuran hidden layer, dan jumlah head attention) untuk mengoptimalkan performa dengan sumber daya yang tersedia.
3. Pendekatan Transfer Learning: Menerapkan teknik transfer learning dari model BERT yang telah dilatih pada bahasa lain, diikuti oleh fine-tuning dengan dataset Bahasa Indonesia.

2.4. Metrik Evaluasi

Evaluasi akan menggunakan metrik yang relevan untuk Bahasa Indonesia dan tugas-tugas NLP yang diuji:

1. Akurasi, Precision, dan Recall: Menggunakan metrik standar ini untuk mengevaluasi performa model dalam berbagai tugas NLP seperti klasifikasi teks, analisis sentimen, dan pemahaman teks.
2. F1 Score: Menyediakan keseimbangan antara precision dan recall, terutama untuk dataset yang tidak seimbang.

3. Analisis Kesalahan: Mendalaminya untuk memahami jenis kesalahan yang dilakukan oleh model dan bagaimana ini berkaitan dengan keunikan Bahasa Indonesia.
4. Penilaian Subjektif: Melibatkan penutur asli Bahasa Indonesia dalam penilaian kualitatif untuk mengevaluasi pemahaman semantik dan nuansa bahasa oleh model.

Dengan pendekatan metodologi ini, penelitian Anda akan secara komprehensif mengevaluasi dan memperbaiki penerapan BERT untuk Bahasa Indonesia, berkontribusi pada pengembangan NLP yang lebih efektif dan inklusif.

3. Hasil dan Pembahasan

3.1. Sumber Dataset

Berikut adalah sumber data terbuka dalam penelitian ini:

1. Indonesian Wikipedia Dump: Data teks dari Wikipedia Bahasa Indonesia. (https://dumps.wikimedia.org/idwiki/latest/)
2. OpenSubtitles: Kumpulan teks subtitle film, termasuk dalam Bahasa Indonesia. Berguna untuk analisis bahasa sehari-hari. (https://www.opensubtitles.org/)
3. PANL10n: Kumpulan data teks dalam berbagai bahasa, termasuk Bahasa Indonesia, dari proyek PAN Localization. (http://www.panl10n.net/)
4. Asian Language Treebank (ALT): Dataset meliputi teks Bahasa Indonesia untuk pelatihan NLP. (https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/)
5. Tatoeba: Koleksi kalimat dan terjemahannya, termasuk Bahasa Indonesia, yang berguna untuk tugas NLP multibahasa. (https://tatoeba.org/)
6. Twitter Indonesian Dataset: Dataset untuk analisis sentimen dan lainnya dari Twitter dalam Bahasa Indonesia. (https://data.mendeley.com/datasets/v524p5dhpj/1)
7. IndoNLU: Kumpulan dataset untuk berbagai tugas NLP dalam Bahasa Indonesia, termasuk klasifikasi teks, analisis sentimen, dan lainnya.(https://www.indobenchmark.com/)
8. Kaggle's Indonesian News Dataset: Kumpulan berita Bahasa Indonesia yang berguna untuk tugas NLP seperti ringkasan teks dan klasifikasi. (https://www.kaggle.com/antoreepjana/indonesian-news-dataset)
9. TED Talks Bahasa Indonesia: Terjemahan transkrip TED Talks ke dalam Bahasa Indonesia, berguna untuk analisis linguistik dan pembelajaran mesin. (https://www.ted.com/participate/translate)

3.2. Eksperimen

***Tabel 1**. Tugas, Metode Dasar dan Metode Fine-Tuning BERT dari berbagai sumber.*

Tugas	Metode Dasar	Model Fine-tuning
POS Tagging dan NER	- Lampe et al. (2016)26: BiLSTM + CRF dengan embedding karakter dan kata (diperbaharui). lr: 0.001, epoch: 100 dengan early stopping (patience = 5)	- Fine-tuning: menambahkan lapisan klasifikasi untuk setiap token, lr: 5e-5, epoch:100 dengan early stopping (patience = 5)
Dependency parsing	1. Dozat dan Manning (2017), Parser Bi-Affine, Embedding: fastText (tetap) 2. Rahman dan Purwarianti (2020)i 3. Kondratyuk dan Straka (2019)† 4. Alnafis et al. (2019)†	Dozat dan Manning (2017), Bi-Affine parser, Embedding: BERT output (tetap)
Analisis Sentimen	1. 200-d BiLSTM Embedding: fastText (tetap), lr: 0.001, epoch: 100 dengan early stopping (patience = 5) 2. Naive Bayes dan	- Fine-tuning: input: 200 token; epoch: 20; lr: 5e-5; batch size: 30; warm-up: 10% dari total steps; early stopping

	Logistic Regression input: Byte-pair encoding (unigram+bigram)27	(patience = 5); Output layer menggunakan encoded [CLS]
Ringkasan	1. Kurniawan dan Louvan (2018)† 2. Cheng dan Lapata (2016)†	Liu dan Lapata (2019), model ekstraktif, 20,000 steps, lr: 2e-3, dan token: 512.28
NTP	200-d BiLSTM (binary-class). Embedding: fastText (tetap), lr: 0.001, epoch:100 dengan early stopping (patience = 20)	- Fine-tuning: input: 60 token (untuk 1 single tweet); epoch: 20; learning rate: 5e-5; batch size: 20; warm-up: 10% dari total steps; early stopping (patience = 5); Output layer menggunakan encoded [CLS]
Urutan Tweet	Hierarchical 200-d BiLSTMs (multi-class). Embedding: fastText (tetap), lr: 0.001, epoch:100 dengan early stopping (patience = 20)	- Fine-tuning: input: 50 token (untuk 1 single tweet); epoch: 20; learning rate: 5e-5; batch size: 20; warm-up: 10% dari total steps; early stopping (patience = 5); BERT fine-tuning disertai dengan trik Liu dan Lapata (2019) (alternated seq.)

Catatan:

"lr" adalah singkatan dari "learning rate".
"epoch" adalah satu siklus pelatihan lengkap pada dataset.
"early stopping" adalah metode untuk menghentikan pelatihan jika performa model tidak meningkat lagi setelah beberapa epoch.
"[CLS]" adalah token khusus yang digunakan di BERT untuk representasi agregat dari sebuah kalimat atau pasangan kalimat.
"patience" adalah jumlah epoch yang diizinkan tanpa peningkatan pada metrik yang dipantau sebelum pelatihan dihentikan.

Untuk membandingkan INDOBERT, kami melakukan perbandingan terhadap dua model BERT yang telah ada sebelumnya: BERT multibahasa ("MBERT"), dan BERT monolingual untuk Bahasa Melayu ("MALAYBERT"). MBERT dilatih dengan cara menggabungkan dokumen Wikipedia untuk 104 bahasa termasuk Bahasa Indonesia, dan telah terbukti efektif untuk tugas-tugas multibahasa zero-shot (Wu dan Dredze, 2019; Wang et al., 2019c). MALAYBERT adalah model yang tersedia untuk umum dan dilatih pada dokumen Bahasa Melayu dari Wikipedia, sumber berita lokal, media sosial, dan beberapa terjemahan dari Bahasa Inggris. Kami berharap MALAYBERT akan menyediakan representasi yang lebih baik daripada MBERT untuk Bahasa Indonesia, karena Bahasa Melayu dan Indonesia saling dapat dimengerti, dengan banyak kemiripan leksikal, namun perbedaan yang nyata dalam tata bahasa, pengucapan, dan kosa kata.

Untuk tugas-tugas pelabelan sekuensial (penandaan POS dan NER), analisis sentimen, NTP, dan tugas pengurutan tweet, prosedur fine-tuning dijelaskan secara rinci dalam Tabel 1.

Untuk parsing dependency, kami mengikuti Nguyen dan Nguyen (2020) dalam menggabungkan BERT ke dalam parser dependensi BiAfine (Dozat dan Manning, 2017) dengan menggantikan embedding kata dengan representasi kontekstual yang sesuai. Secara khusus, kami menghasilkan embedding BERT dari token WordPiece pertama sebagai embedding kata, dan melatih parser BiAfine dalam konfigurasi defaultnya. Selain itu, kami juga melakukan benchmark terhadap versi MBERT yang telah fine-tuned sebelumnya yang dilatih lebih dari 75 kumpulan data UD yang digabungkan dalam berbagai bahasa (Kondratyuk dan Straka, 2019).

Untuk ringkasan, kami mengikuti Liu dan Lapata (2019) dalam mengkode dokumen dengan memasukkan token [CLS] dan [SEP] antar kalimat. Kami juga menerapkan embedding segmen bergantian berdasarkan

apakah posisi sebuah kalimat ganjil atau genap. Di atas model pra-latih, kami menggunakan encoder transformer kedua untuk mempelajari hubungan antar kalimat. Inputnya adalah representasi [CLS] yang telah dikodekan, dan outputnya adalah label ekstraktif $y \in \{0, 1\}$ (1 = termasuk dalam ringkasan; 0 = tidak termasuk).

3.3. Hasil dan Pembahasan

***Tabel 2**. Tabel Hasil Perbandingan Berbagai Metode BERT*

Metode	POS Tagging Acc	NER UGM F1	NER UI F1
BiLSTM-CRF (Lample et al., 2016)	95.4	70.9	82.2
MBERT	96.8	71.6	82.2
MALAYBERT	96.8	73.2	87.4
INDOBERT	**96.8**	**74.9**	**90.1**

Catatan:
"Acc" adalah singkatan dari "Accuracy" atau "Akurasi".
"F1" merujuk pada skor F1, yang merupakan ukuran yang digunakan untuk menguji akurasi dari model klasifikasi.
"POS Tagging" merujuk pada proses penentuan bagian dari ucapan kata dalam teks.
"NER" adalah singkatan dari "Named Entity Recognition", yang merujuk pada proses mengidentifikasi dan klasifikasi elemen penting dalam teks menjadi kategori yang telah ditentukan seperti nama entitas, lokasi, dll.
"UGM" dan "UI" dataset yang digunakan untuk menguji model tersebut (berasal dari UI dan UGM).

Hasil dalam Tabel 2 membandingkan berbagai metode BERT untuk Akurasi Tagging POS dan skor F1 Named Entity Recognition (NER) menggunakan dataset UGM dan UI. Perbandingan mencakup metode BiLSTM-CRF, MBERT, MALAYBERT, dan INDOBERT.

Akurasi Tagging POS: Mengukur efektivitas masing-masing metode dalam mengidentifikasi bagian ucapan di teks. Semua metode berbasis BERT (MBERT, MALAYBERT, INDOBERT) menunjukkan akurasi tinggi yang serupa sebesar 96.8%, mengungguli metode BiLSTM-CRF. Skor F1 NER UGM dan UI: Skor ini menilai presisi dan recall dalam mengidentifikasi dan mengklasifikasi elemen penting seperti nama entitas dan lokasi dalam teks. Dalam skor F1 NER, INDOBERT mengungguli metode lain secara signifikan pada dataset UGM dan UI, menunjukkan kemampuannya yang lebih unggul dalam mengenali entitas bernama dalam teks bahasa Indonesia. MALAYBERT juga menunjukkan performa yang kuat, terutama pada dataset UI.

Hasil ini menunjukkan bahwa INDOBERT dan MALAYBERT, yang lebih spesifik untuk bahasa Indonesia dan Melayu, memiliki keunggulan dibandingkan MBERT yang lebih umum dan metode non-transformer BiLSTM-CRF dalam tugas NLP spesifik ini. Ini menandakan pentingnya penyesuaian model spesifik bahasa dalam mencapai kinerja yang lebih tinggi dalam aplikasi NLP.

4. Kesimpulan

Penelitian ini berhasil menunjukkan bahwa model BERT dapat diadaptasi secara efektif untuk Bahasa Indonesia, meskipun ada tantangan unik yang terkait dengan struktur linguistik dan variasi dialek bahasa ini. Melalui eksperimen yang komprehensif dan analisis mendalam, penelitian ini membuktikan bahwa BERT, saat disesuaikan dan dioptimalkan dengan dataset yang representatif dan berkualitas, dapat meningkatkan pemrosesan bahasa alami untuk Bahasa Indonesia. Hal ini menggarisbawahi pentingnya adaptasi model NLP berbasis transformer untuk berbagai bahasa, serta menyediakan wawasan untuk peningkatan teknologi NLP yang lebih inklusif dan efektif di masa depan.

5. Referensi

Bai, J., Wang, Y., Chen, Y., Yang, Y., Bai, J., Yu, J., & Tong, Y. (2021). Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees.

Chaudhry, P. (2022). Bidirectional Encoder Representations from Transformers for Modelling Stock Prices. International Journal for Research in Applied Science and Engineering Technology.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2019). Pre-Training With Whole Word Masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3504-3514.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting Pre-Trained Models for Chinese Natural Language Processing. ArXiv.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT Model. ArXiv.

Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2020). ParsBERT: Transformer-based Model for Persian Language Understanding. Neural Processing Letters, 53, 3831-3847.

Lee, E., Lee, C., & Ahn, S. (2022). Comparative Study of Multiclass Text Classification in Research Proposals Using Pre Trained Language Models. Applied Sciences.

Min, B., Ross, H. H., Sulem, E., Ben Veyseh, A. P., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2021). Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. ArXiv.

Nityasya, M. N., Wibowo, H. A., Chevi, R., Prasojo, R. E., & Aji, A. F. (2022). Which Student is Best? A Comprehensive Knowledge Distillation Exam for Task-Specific BERT Models. ArXiv.

Nozza, D., Bianchi, F., & Hovy, D. (2020). What the [MASK]? Making Sense of Language-Specific BERT Models. ArXiv.

Paass, G., & Giesselbach, S. (2023). Foundation Models for Natural Language Processing - Pre-trained Language Models Integrating Media. ArXiv.

Perez, I., & Reinauer, R. (2022). The Topological BERT: Transforming Attention into Topology for Natural Language Processing. ArXiv.

Sebastian, D., Purnomo, H., & Sembiring, I. (2022). BERT for Natural Language Processing in Bahasa Indonesia. 2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), 204-209.

Wijayanti, R. (2023). Transfer learning berbasis cross-lingual word embedding untuk peringkasan teks ekstraktif. Institut Teknologi Bandung. https://repository.itb.ac.id/thesis/index.php.

Zhang, X., Ding, Y., Yu, M., O'Uchi, S., & Fujita, M. (2022). Low-Precision Quantization Techniques for Hardware-Implementation-Friendly BERT Models. 2022 23rd International Symposium on Quality Electronic Design (ISQED).