PENGARUH AUGMENTASI DATA TERHADAP KINERJA *PRE-TRAINED TRANSFORMER MODELS* DALAM KLASIFIKASI JUDUL BERITA COVID-19

(Skripsi)

Oleh

YULINA PUTRI
NPM. 2117031084

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRACT

THE EFFECT OF DATA AUGMENTATION ON THE PERFORMANCE OF PRE-TRAINED TRANSFORMER MODELS IN COVID-19 NEWS TITLE CLASSIFICATION

By

YULINA PUTRI

The volume of digital information, especially online news during the COVID-19 pandemic has surged, making it difficult for readers to obtain relevant information. Automatic classification based on pre-trained transformer models is an approach that can be used to filter information efficiently. However, the performance of classification models is greatly affected by the quality and distribution of training data. The imbalance of class distribution in the data can degrade the model's prediction performance, especially for minority classes. Therefore, this study aims to analyze the effect of data augmentation on improving the performance of classification models on imbalanced COVID-19 news title data. Two augmentation methods are used, namely synonym replacement (SR) and a combination of synonym replacement with back translation (SR + BT). The evaluation was carried out using several pre-trained transformer models, including ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, and RoBERTa. The results show that the augmentation successfully improved the average accuracy and recall up to 1.45%, precision and F1-score up to 1.94%, and ROC-AUC up to 5.58%. In the minority class, augmentation increased recall by 14.67% and F1-score by 5.37%. This finding shows that data augmentation is effective in overcoming class imbalance and improving the model's ability to identify minority classes that were previously difficult to detect.

Keywords: Classification, Transformer, Data Augmentation, Online News, Synonym Replacement, Back Translation.

ABSTRAK

PENGARUH AUGMENTASI DATA TERHADAP KINERJA *PRE-TRAINED TRANSFORMER MODELS* DALAM KLASIFIKASI JUDUL BERITA COVID-19

Oleh

YULINA PUTRI

Volume informasi digital, khususnya berita *online* selama masa pandemi COVID-19 mengalami lonjakan yang menyebabkan pembaca kesulitan memperoleh informasi yang relevan. Klasifikasi otomatis berbasis *pre-trained transformer models* merupakan pendekatan yang dapat digunakan untuk menyaring informasi secara efisien. Namun, kinerja model klasifikasi sangat dipengaruhi oleh kualitas dan distribusi data latih. Ketidakseimbangan distribusi kelas dalam data dapat menurunkan kinerja prediksi model, terutama terhadap kelas minoritas. Oleh karena itu, penelitian ini bertujuan menganalisis pengaruh augmentasi data terhadap peningkatan kinerja model klasifikasi pada data judul berita COVID-19 yang tidak seimbang. Dua metode augmentasi yang digunakan yaitu *synonym replacement* (SR) dan kombinasi *synonym replacement* dengan *back translation* (SR + BT). Evaluasi dilakukan menggunakan beberapa model *pre-trained transformer*, diantaranya ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, dan RoBERTa. Hasil menunjukkan bahwa augmentasi berhasil meningkatkan rata-rata akurasi dan *recall* hingga 1,45%, presisi dan *F1-score* hingga 1,94%, dan ROC-AUC hingga 5,58%. Pada kelas minoritas, augmentasi meningkatkan *recall* hingga 14,67% dan *F1-score* hingga 5,37%. Temuan ini menunjukkan bahwa augmentasi data efektif dalam mengatasi ketidakseimbangan kelas dan meningkatkan kemampuan model dalam mengidentifikasi kelas minoritas yang sebelumnya sulit terdeteksi.

Kata-kata kunci: Klasifikasi, *Transformer*, Augmentasi Data, Berita *Online*, *Synonym Replacement*, *Back Translation*.

PENGARUH AUGMENTASI DATA TERHADAP KINERJA *PRE-TRAINED TRANSFORMER MODELS* DALAM KLASIFIKASI JUDUL BERITA COVID-19

YULINA PUTRI

(Skripsi)

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika

Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

Judul Skripsi	:	**PENGARUH AUGMENTASI DATA TERHADAP KINERJA *PRE-TRAINED TRANSFORMER MODELS* DALAM KLASIFIKASI JUDUL BERITA COVID-19**
Nama Mahasiswa	:	**Yulina Putri**
Nomor Pokok Mahasiswa	:	**2117031084**
Program Studi	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Dian Kurniasari, S.Si., M.Sc.
NIP. 196903051996032001

Dr. Purnomo Husnul Khotimah, M.T.
NIP. 198003232005022002

2. Ketua Jurusan Matematika

Dr.Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Dian Kurniasari, S.Si., M.Sc.** .....

Sekretaris : **Dr. Purnomo Husnul Khotimah, M.T.** .....

Penguji
Bukan Pembimbing : **Ir. Warsono, M.S., Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: **10 Juni 2025**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	:	**Yulina Putri**
Nomor Pokok Mahasiswa	:	**2117031084**
Jurusan	:	**Matematika**
Judul Skripsi	:	**Pengaruh Augmentasi Data Terhadap Kinerja *Pre-Trained Transformer Models* Dalam Klasifikasi Judul Berita COVID-19**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri. Apabila kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 10 Juni 2025

Penulis

Yulina Putri

RIWAYAT HIDUP

Penulis bernama lengkap Yulina Putri, lahir di Lampung, 21 Desember 2003. Penulis merupakan putri sulung dari Bapak Syamsudin dan Ibu Noviyanti Yuniar, memiliki seorang adik yang bernama Meila Syafitri.

Penulis menempuh pendidikan di TK AL HIKMAH pada tahun 2008-2009. Pendidikan Dasar di SD Negeri 4 Merapi Barat pada tahun 2009-2015. Pendidikan Menengah Pertama di SMP Negeri 2 Merapi Barat pada tahun 2015-2018. Pendidikan Menengah atas di SMA Negeri 1 Merapi Barat pada tahun 2018-2021.

Pada tahun 2021, penulis melanjutkan studi ke jenjang Perguruan Tinggi di Program Studi S1 Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung melalui Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN). Selama menjadi mahasiswi, penulis aktif dalam kegiatan organisasi, kepanitiaan, dan berbagai program lainnya. Pada tahun 2022 penulis menjadi pengurus ROIS FMIPA Unila sebagai anggota Biro Dana dan Usaha. Pada tahun 2023, penulis melanjutkan kepengurusan di ROIS FMIPA Unila dan menjabat sebagai Bendahara Umum.

Pada tahun 2024, penulis menjalankan Praktik Kerja Lapangan (PKL) dan Merdeka Belajar Kampus Merdeka (MBKM) Penelitian/Riset di Badan Riset dan Inovasi Nasional (BRIN) KST Samaun Sadikun Bandung, Jawa Barat. Pada tahun yang sama, penulis mengikuti kegiatan Studi Independen Kampus Merdeka di PT Revolusi Cita Edukasi. Pada tahun ajaran semester ganjil 2024/2025, penulis diamanahkan sebagai asisten dosen mata kuliah Eksplorasi Data.

KATA INSPIRASI

”Maka sesungguhnya bersama kesulitan ada kemudahan, sesungguhnya bersama kesulitan ada kemudahan.”
(QS. Al-Insyirah: 5-6)

Katakanlah, ”Sesungguhnya salatku, ibadahku, hidupku, dan matiku hanyalah untuk Allah, Tuhan semesta alam.”
(QS. Al-An’am: 162)

”Maka bersabarlah sesungguhnya janji Allah itu benar.”
(QS. Ar-Rum: 60)

Ketika semua terasa mustahil, yakinlah ada Allah yang dapat membuat semuanya menjadi mungkin karena ”innama amruhu idza arada syai’an ay yaqula lahu kun fa yakun” (QS.36:82)

”Tantangan terbesar dalam hidupmu adalah dirimu sendiri, semua perubahan datang dari dirimu, untuk dirimu”

”Pursue the hereafter and the world will follow”

”InsyaAllah lelah kan jadi lillah”

PERSEMBAHAN

Bismillahirrahmanirrahim, alhamdulillahirabbil 'aalamiin, puji dan syukur kepada Allah Swt. atas segala nikmat serta hidayah-Nya sehingga skripsi ini dapat diselesaikan dengan baik. Salawat serta salam semoga selalu tercurahkan kepada junjungan Nabi Muhammad saw. yang telah menjadi suri tauladan bagi umat manusia. Dengan rasa syukur, penulis mempersembahkan rasa terima kasih kepada:

Orang Tuaku Tercinta

Terima kasih atas segala doa, kasih sayang, pengorbanan, motivasi, dan dukungan tanpa henti. Karya ini adalah persembahan kecil untuk cinta dan pengorbanan yang begitu besar. Semoga Allah senantiasa menjaga dan menyayangi mamak & bapak, sebagaimana mamak & bapak menjaga dan menyayangiku.

Dosen Pembimbing dan Pembahas

Terima kasih kepada Ibu/Bapak atas waktu, ilmu, motivasi, serta arahan yang telah berperan penting dalam penyusunan karya ini dan pembentukan sikap ilmiah sepanjang proses akademik. Semoga Allah Swt. senantiasa melimpahkan berkah dan rahmat-Nya.

Keluarga Besar

Terima kasih telah memberikan dukungan, nasihat, dan doa yang mengiringi setiap langkah.

Teman-temanku

Terima kasih telah menjadi teman cerita, belajar, dan bertumbuh. Semoga langkah kita seiring dalam kebaikan, menuju keberhasilan dan keberkahan dunia akhirat.

Almamater Tercinta

Universitas Lampung

SANWACANA

Alhamdulillah, puji dan syukur penulis panjatkan kepada Allah Swt. atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan judul **"Pengaruh Augmentasi Data Terhadap Kinerja *Pre-Trained Transformer Models* Dalam Klasifikasi Judul Berita COVID-19"** dengan baik dan lancar serta tepat pada waktu yang telah ditentukan. Salawat serta salam semoga senantiasa tercurahkan kepada Nabi Muhammad saw. Banyak pihak yang telah membantu memberikan dukungan, arahan, motivasi serta saran sehingga skripsi ini dapat terselesaikan. Oleh karena itu, dalam kesempatan ini penulis mengucapkan terima kasih kepada:

1. Ibu Dr. Dian Kurniasari, S.Si., M.Sc. selaku dosen pembimbing I yang telah banyak meluangkan waktunya untuk memberikan arahan, bimbingan, bantuan, motivasi, saran serta dukungan kepada penulis dalam menyusun skripsi ini.
2. Ibu Dr. Purnomo Husnul Khotimah, M.T. selaku pembimbing II yang telah banyak meluangkan waktunya untuk memberikan arahan, bimbingan, bantuan, motivasi, saran serta dukungan kepada penulis dalam menyusun skripsi ini.
3. Bapak Ir. Warsono, M.S., Ph.D. selaku dosen penguji yang telah bersedia memberikan evaluasi, kritik, serta saran yang membangun selama proses penyusunan skripsi.
4. Bapak Andri Fachrur Rozie, S.Kom., M.Eng. selaku salah satu pembimbing MBKM BRIN KST Samaun Sadikun, Bandung.
5. Ibu Dr. Notiragayu, S.Si., M.Si. selaku dosen pembimbing akademik.
6. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.
7. Seluruh dosen, staff dan civitas akademika Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.
8. Kedua orangtua yang sangat penulis sayangi, Bapak Syamsudin dan Ibu Noviyanti Yuniar yang senantiasa mendoakan, menyayangi, mendidik, membimbing, memotivasi dan mendukung penulis dalam setiap perjalanan.
9. Adik tersayang Meila Syafitri yang telah mendoakan dan mendukung penulis.

10. Keluarga besar, guru, dan semua orang baik yang telah memberikan doa, semangat, dan dukungan kepada penulis.
11. Folandia, yang telah penulis anggap seperti saudari sendiri, terima kasih atas doa, dukungan, semangat, dan kebersamaannya.
12. Anastasia, Clarisa, Cantika, Rani, Dinda, Mia, rekan seperjuangan kuliah. Terima kasih untuk waktu, dukungan, dan cerita yang telah kita lewati.
13. Seluruh peneliti, staff, karyawan, dan rekan MBKM BRIN KST Samaun Sadikun Bandung, Rois FMIPA Unila Periode 2023 yang telah memberikan inspirasi dan pengalaman hidup.
14. Semua pihak yang telah terlibat dalam penyusunan skripsi ini, yang tidak dapat penulis sebutkan satu per satu.

Semoga skripsi ini dapat bermanfaat bagi kita semua. Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan, sehingga penulis mengharapkan kritik dan saran yang membangun untuk menjadikan skripsi ini lebih baik lagi.

Bandar Lampung, 10 Juni 2025

Yulina Putri

DAFTAR ISI

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Pada kehidupan serba digital, informasi telah menjadi bagian yang tidak terpisahkan dari aktivitas manusia saat ini. Namun, semakin berkembangnya teknologi mengakibatkan munculnya masalah lonjakan volume informasi yang besar, yang akan terus meningkat dari waktu ke waktu. Penemuan mesin cetak oleh Gutenberg sekitar tahun 1540 menandai perubahan besar dalam sejarah penyebaran informasi (Benselin dan Ragsdel, 2016). Seiring kemajuan teknologi, media seperti radio, televisi, dan komputer mulai digunakan sebagai alat untuk menyebarkan informasi (Khan dkk., 2020). Pada akhir abad ke-20, internet muncul sebagai sarana baru untuk mengakses informasi secara global yang dapat dijangkau dengan mudah tanpa terikat waktu maupun lokasi. Namun, kemudahan ini tidak selalu membawa dampak positif. Kemudahan ini justru mengakibatkan jumlah permintaan informasi terus meningkat bahkan bisa melampaui batas kapasitas yang tersedia, hal ini dapat menyebabkan kebingungan pengguna dalam memproses informasi yang tersedia (Shahrzadi dkk., 2024).

Lonjakan informasi mengalami peningkatan yang signifikan, terutama selama masa pandemi COVID-19 yang terjadi pada tahun 2019 lalu. Pada saat itu, kondisi sangat menghawatirkan, hari demi hari pandemi semakin meluas menyebabkan jutaan jiwa kehilangan nyawa, ekonomi menurun, jalanan menjadi sepi, sekolah dan kantor ditutup, semua orang diimbau untuk tetap di rumah. Banyak negara termasuk Indonesia menerapkan berbagai pembatasan untuk menekan penyebaran virus. Masyarakat menjalani karantina di rumah dan dilarang melakukan kegiatan yang melibatkan kerumunan guna mencegah penyebaran COVID-19. Ketakutan, kebingungan, dan kecemasan menjadi bagian dari kehidupan sehari-hari. Semua kebiasaan seperti berkumpul, berjabat tangan, atau bepergian dianggap berbahaya dan menjadi hal yang berisiko. Hal ini menyebabkan kepanikan masyarakat

dalam menghadapi situasi yang semakin sulit. Pada kondisi ini, masyarakat sangat membutuhkan informasi akurat dan terkini seperti jumlah kasus terkonfirmasi dan langkah pencegahan pandemi. Oleh sebab itu, masyarakat secara aktif menggunakan internet dan media *online* untuk memperoleh informasi terkini mengenai situasi dan perkembangan COVID-19 (Zheng dkk., 2023).

Tingginya kewaspadaan terhadap COVID-19 menyebabkan konsumsi informasi daring terutama dalam bentuk berita *online* meningkat drastis. Setiap hari, berbagai portal berita *online* di Indonesia menerbitkan ribuan judul artikel dengan beragam topik dan format penulisan (Krstajic dkk., 2013). Meningkatnya popularitas berita *online*, mengakibatkan munculnya berbagai tantangan dalam mengelola, mengklasifikasikan, dan mencari informasi yang relevan bagi pembaca. Namun, manusia memiliki keterbatasan untuk melakukan klasifikasi secara manual, sehingga diperlukan sistem yang dapat mengklasifikasikan dan mengorganisir informasi secara otomatis untuk mempermudah proses penyaringan dan pencarian berita yang relevan.

Pembelajaran mendalam menjadi bagian kecerdasan buatan yang telah mengalami perkembangan pesat dalam beberapa dekade terakhir. Berbagai model pembelajaran mendalam seperti jaringan saraf tiruan telah banyak digunakan untuk tugas klasifikasi. *Convolutional Neural Network* (CNN) sangat baik dalam mengenali pola spasial dalam data seperti gambar (Dai dkk., 2024), sedangkan *Recurrent Neural Network* (RNN) lebih cocok untuk data sekuensial seperti teks (Elbasani dkk., 2021). Meskipun CNN dan RNN mencapai banyak keberhasilan, model-model ini memiliki keterbatasan dalam menangkap hubungan antar data yang panjang. *Transformer* muncul sebagai alternatif. Meskipun tergolong model bahasa terbaru, *transformer* memiliki kompleksitas yang lebih tinggi dan unggul dalam menangkap dependensi jarak jauh serta mampu mengidentifikasi hubungan antar kata dalam teks. Model ini juga efektif dalam memahami makna dalam teks panjang, seperti judul berita yang sering kali mengandung frasa padat makna atau istilah yang memerlukan pemahaman mendalam. Selain itu, *transformer* telah dirancang khusus untuk tugas-tugas pemrosesan bahasa alami dan terbukti sangat efektif untuk mengatasi berbagai tugas lainnya, termasuk klasifikasi.

Beberapa keberhasilan *transformer* dalam tugas klasifikasi diantaranya Khotimah dkk. (2023) menemukan bahwa model *Indonesian BERT* (IndoBERT) dengan *fine tuning* menghasilkan kinerja yang lebih unggul dibandingkan model *machine*

learning dan *deep learning* berbasis jaringan saraf tiruan dalam mengklasifikasikan data judul berita *online*. Gonzalez-Carvajal dan Garrido-Merchan (2020) menemukan bahwa *Bidirectional Encoder Representations from Transformers* (BERT) lebih efektif dibandingkan dengan berbagai metode pembelajaran mesin tradisional dalam tugas klasifikasi teks dengan jenis dataset yang berbeda-beda. Model *transformer* seperti *Robustly Optimized BERT Approach* (RoBERTa) lebih unggul dibandingkan model BERT, *A Lite BERT* (ALBERT), dan *Distilled BERT* (DistilBERT) dalam mendeteksi ulasan palsu (Gupta dkk., 2021). Model IndoBERT menghasilkan kinerja lebih tinggi dibandingkan BERT dan RoBERTa dalam kasus klasifikasi *headline clickbait* Indonesia dengan akurasi sebesar 92.41%. (Sirusstara dkk., 2022). Selain itu, Nabiilah dkk. (2023) menemukan bahwa *pre-trained transformer model* yang dilatih khusus untuk bahasa Indonesia, seperti IndoBERT, terbukti lebih efektif dalam klasifikasi komentar *toxic* dibandingkan model *transformer* lainnya.

Meskipun *transformer* efektif dalam berbagai tugas klasifikasi dan mampu menangkap hubungan antar kata yang kompleks, kinerja klasifikasi dengan model *transformer* sangat bergantung pada ukuran dan kualitas kumpulan data pelatihan. Kinerja yang buruk akan terjadi jika kumpulan data yang tidak mencukupi dan tidak seimbang (Sharifirad dkk., 2018). Ketidakseimbangan kelas adalah masalah umum untuk banyak tugas klasifikasi pemrosesan bahasa alami. Ketidakseimbangan kelas terjadi ketika salah satu kategori kelas memiliki bagian yang jauh lebih besar daripada kategori kelas lainnya. Hal ini mengakibatkan proses klasifikasi menjadi lebih sulit untuk mengidentifikasi kelas minoritas (Madabushi dkk., 2020). Selain itu, klasifikasi data tidak seimbang dengan variabel target yang tidak merata akan menyebabkan bias terhadap kelas mayoritas dalam prediksi (Vairetti dkk., 2024). Oleh karena itu, menyeimbangkan data yang tidak seimbang sangat penting.

Cochran dkk. (2022) menemukan bahwa efek dari set data yang seimbang dan jumlah data yang lebih besar dapat meningkatkan akurasi klasifikasi. Berbagai teknik dapat diterapkan untuk mengatasi masalah ketidakseimbangan dataset. Salah satu pendekatannya adalah *undersampling*, dimana jumlah data pada kelas mayoritas diturunkan sehingga teknik ini efektif untuk dataset dengan rasio ketidakseimbangan yang lebih rendah (Devi dkk., 2020), sebaliknya *oversampling* dapat meningkatkan ukuran set pelatihan dengan menambah data kelas minoritas. *Oversampling* dapat mempertahankan data yang ada sehingga informasi penting dari kelas mayoritas tetap tersedia. Teknik *oversampling* seperti augmentasi data

dengan sinonim dan *back translation* menyebabkan model mendapat manfaat dari data pelatihan tambahan yang dihasilkan dan berhasil meningkatkan kinerja model (Rahma dan Suadaa, 2023). Ningsih dkk. (2022) juga menemukan bahwa *synonym text generation* sangat mempengaruhi kinerja model *deep learning* seperti *Multilayer Perceptron* (MLP), *Convolutional Neural Network* (CNN), dan *Long Short-Term Memory* (LSTM), teknik ini mampu meningkatkan kumpulan data pelatihan dan generalisasi model dengan peningkatan akurasi hingga 4%. Wirawan dan Cahyono (2023) juga menyatakan bahwa teknik augmentasi terbaik diperoleh dengan mengubah kata-kata menjadi sinonim.

Berdasarkan uraian diatas, penelitian ini bertujuan mengatasi ketidakseimbangan dataset judul berita *online* berbahasa Indonesia dengan melakukan *oversampling* yaitu melalui augmentasi data latih menggunakan penggantian kata dengan sinonim atau kata yang memiliki makna sama. Selain itu, *back translation* juga digunakan untuk menambah variasi data latih dengan menghasilkan kalimat baru. Kemudian, hasil data seimbang dan tidak seimbang akan diklasifikasikan menggunakan berbagai model *transformer* yang telah dilatih sebelumnya atau disebut *pre-trained transformer models*. Selanjutnya, melakukan *fine-tuning* untuk mengoptimalkan kinerja model. Terakhir, kinerja model akan dievaluasi menggunakan metrik akurasi, presisi, *recall*, *F1-score*, dan *Receiver Operating Characteristic – Area Under the Curve* (ROC-AUC). Hasil evaluasi model menggunakan data seimbang dan data tidak seimbang tersebut akan dibandingkan.

Penelitian ini disusun menjadi 5 bagian. Bagian 1 menyajikan latar belakang, rumusan masalah, tujuan penelitian, dan manfaat penelitian. Bagian 2 berisi kajian terdahulu dan literatur yang berkaitan dengan penelitian. Bagian 3 menyajikan metode dan kerangka kerja penelitian. Bagian 4 membahas dan memberikan analisis mendalam terhadap hasil penelitian. Terakhir, bagian 5 menyatakan kesimpulan penelitian.

1.2 Rumusan Masalah

Adapun rumusan masalah dalam penelitian ini yaitu:

1. Menganalisis kinerja berbagai *pre-trained transformer models*, seperti ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, dan RoBERTa dalam melakukan klasifikasi judul berita COVID-19.

2. Mengaplikasikan metode augmentasi *synonym replacement* dan augmentasi *synonym replacement* yang dikombinasikan dengan *back translation* terhadap klasifikasi data judul berita COVID-19 yang memiliki distribusi kelas tidak seimbang.

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini yaitu:

1. Membandingkan hasil kinerja berbagai *pre-trained transformer models*, seperti ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, dan RoBERTa dalam klasifikasi judul berita COVID-19.
2. Mengetahui pengaruh augmentasi *synonym replacement* dan augmentasi *synonym replacement* yang dikombinasikan dengan *back translation* dalam mengklasifikasikan judul berita COVID-19 berdasarkan berbagai *pre-trained transformer models*, diantaranya ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, dan RoBERTa

1.4 Manfaat Penelitian

Adapun manfaat penelitian ini diantaranya:

1. Memberikan edukasi terkait pengembangan ilmu matematika dalam menangani ketidakseimbangan dataset serta dapat menjadi bahan informasi tambahan khususnya dalam penerapan pemrosesan bahasa alami di bidang *data mining* dengan metode *deep learning* yaitu *pre-trained transformer models*.
2. Memberikan wawasan keilmuan bahwasanya augmentasi data dapat mempengaruhi kinerja *pre-trained transformer models* untuk klasifikasi teks.
3. Sebagai sistem klasifikasi teks otomatis yang berguna untuk mempercepat pengolahan informasi relevan di masa mendatang.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Sub bagian ini menguraikan beberapa penelitian yang telah dilakukan oleh peneliti sebelumya dan digunakan sebagai referensi untuk penelitian ini. Ringkasan penelitian terdahulu disajikan dalam Tabel 1. sebagai berikut:

Tabel 1. Penelitian Terdahulu

No	**Penelitian**	**Data**	**Metode**	**Hasil (%)**			
				Acc	*Prec*	*Rec*	*F1*
1.	*Monitoring Indonesian online News for COVID-19 Event Detection using Deep Learning* (Khotimah dkk., 2023)	Data judul berita *online*.	IndoBERT	95,16	94,71	94,32	94,51
2.	*Synonym-based Text Generation in Restructuring Imbalanced Dataset for Deep Learning Models* (Ningsih dkk., 2022)	Data judul berita *online*.	MLP (Original Dataset)	91,28	89.32	88.35	88.81
			MLP (Balanced Dataset dengan *Synonym Text Generation*)	95,01	95.05	95.02	95.01

No	Penelitian	Data	Metode	Hasil (%)			
				Acc	*Prec*	*Rec*	*F1*
3.	Penerapan *Text Augmentation* Untuk Mengatasi Data Yang Tidak Seimbang Pada Klasifikasi Teks Berbahasa Indonesia Studi Kasus: Deteksi Judul *Clickbait* dan Komentar *Hate Speech* Pada Berita *online* (Rahma dan Suadaa, 2023)	Dataset *Clickbait*	IndoBERT Original	-	-	-	83,52
			Synonym Replacement	-	-	-	83,23
			Back Translation	-	-	-	84,04
		Dataset *Hate Speech*	IndoBERT Original	-	-	-	61,79
			Synonym Replacement	-	-	-	64,50
			Back Translation	-	-	-	66,97
4.	*Comparing BERT against traditional machine learning text classification* (Gonzalez-Carvajal dan Garrido-Merchan, 2020)	IMDB dataset	BERT	93.87	-	-	-
		Real Or Not Tweet		83,61	-	-	-
		Berita Portugis		90,93	-	-	-
		Ulasan hotel di China.		93,81	-	-	-

No	Penelitian	Data	Metode	Hasil (%)			
				Acc	*Prec*	*Rec*	*F1*
5.	*Clickbait Headline Detection in Indonesian News Sites using Robustly Optimized BERT Pre-training Approach (RoBERTa)* (Sirusstara dkk., 2022)	Dataset CLICK-ID.	IndoBERT	-	92,41	92,23	92,32

2.1.1 Penelitian Pertama (Khotimah dkk., 2023)

Penelitian Khotimah dkk. (2023) berfokus pada pemantauan berita *online* Indonesia untuk deteksi peristiwa COVID-19 menggunakan metode *machine learning* dan *deep learning*. Data dikumpulkan dengan metode *crawling* dari berbagai portal berita *online* Indonesia, khususnya judul berita yang berkaitan dengan peristiwa COVID-19. Khotimah dkk. (2023) menggunakan berbagai metode *machine learning* seperti *Naive Bayes, Logistic Regression, Decision Tree, Support Vector Machine, AdaBoost, Neural Network*, dan *deep learning* seperti MLP, CNN, dan LSTM, termasuk *transformer* khususnya IndoBERT. Hasil menunjukkan bahwa IndoBERT dengan *fine-tuning* menghasilkan akurasi 95,16%, presisi 94,71%, *recall* 94,32%, dan *F1-score* 94,51%. Metode ini lebih unggul dibandingkan model *machine learning* dan *deep learning* lainnya. Selain itu, proses *fine-tuning* mampu meningkatkan kinerja model hingga 57,96%. Kerangka kerja yang dihasilkan mampu mendeteksi kasus COVID-19 di Inggris 13 hari lebih awal, menunjukkan potensinya untuk pemantauan kesehatan masyarakat yang tepat waktu.

2.1.2 Penelitian Kedua (Ningsih dkk., 2022)

Pembelajaran mendalam akan menghasilkan kinerja lebih baik dengan data seimbang. Oleh sebab itu, Ningsih dkk. (2022) menggunakan penambahan teks berbasis sinonim yang bersumber dari Kateglo API untuk membuat kumpulan

data seimbang. Prosesnya berupa penggantian kata-kata dalam kalimat dengan sinonimnya untuk membuat variasi data namun tetap mempertahankan makna asli. Data yang digunakan adalah kumpulan judul berita *online* terkait peristiwa COVID-19. Penggunaan data seimbang dengan fitur *Word embedding* pada tiga model pembelajaran mendalam yaitu MLP, CNN, dan LSTM berhasil meningkatkan akurasi sebesar 2-4%. Namun, model dengan *Term Frequency Inverse Document Frequency* (TFIDF) mengalami penurunan kinerja pada data seimbang. LSTM secara konsisten mencapai kinerja tertinggi di berbagai metrik. Pada *Deep Neural Network - Term Frequency Inverse Document Frequency* (DNN-TFIDF), LSTM mencapai akurasi 80%. Sedangkan pada *Deep Neural Network-Word Embedding* (DNN-WE), LSTM mencapai akurasi hingga 93,49% saat menggunakan kumpulan data seimbang. Kinerja ini secara signifikan lebih baik daripada MLP dan CNN, yang menunjukkan penurunan nilai akurasi hingga 9% saat menggunakan DNN-TFIDF.

2.1.3 Penelitian Ketiga (Rahma dan Suadaa, 2023)

Rahma dan Suadaa (2023) menggunakan 2 jenis dataset yaitu dataset *clickbait* dan dataset *hate speech*. Dataset *clickbait* merupakan kumpulan data judul berita *online* berbahasa Indonesia (CLICK-ID). Dataset ini memiliki 2 label yaitu *clickbait* sejumlah 3.316 dan *non clickbait* sejumlah 5.297. Sedangkan dataset *hate speech* berupa komentar berita *online* berbahasa Indonesia yang memiliki label *hate* sejumlah 263 dan *no hate* sejumlah 1.307. Kedua dataset ini memiliki perbandingan jumlah label yang cukup tinggi dan tergolong dalam kategori dataset yang tidak seimbang. Oleh karena itu, Rahma dan Suadaa (2023) melakukan augmentasi teks melalui 2 metode yaitu *synonym replacement* dan *back translation*. *Synonym replacement* dilakukan dengan mengganti kata dengan sinonim berdasarkan sinonim dari Tesaurus. *Back translation* dilakukan dengan menerjemahkan teks asli kelima bahasa berbeda yaitu Inggris, Cina, Melayu, Jawa, dan tagalog kemudian hasilnya diterjemahkan kembali ke bahasa Indonesia menggunakan API *Google Translate*.

Model klasifikasi yang digunakan yaitu SVM dan IndoBERT. Hasil menunjukkan bahwa teknik *back translation* memberikan peningkatan *F1-score* yang lebih besar dibanding *synonym replacement*. Pada dataset *clickbait*, *back translation* mampu meningkatkan *F1-score* hingga 0,06% pada model SVM dan meningkat 0,53%

pada model IndoBERT. Sedangkan *synonym replacement* mengalami penurunan *F1-score* hingga 0,09% pada model SVM dan menurun 0,29% pada model IndoBERT. Pada dataset *hate speech*, *back translation* mampu meningkatkan *F1-score* hingga 3,72% pada model SVM dan meningkat 5,19% pada model IndoBERT. *Synonym replacement* mampu meningkatkan *F1-score* hingga 3,23% pada model SVM dan meningkat 2,71% pada model IndoBERT. Metode *back translation* menghasilkan kinerja terbaik dengan *F1-score* tertinggi sebesar 84,04% dengan pada IndoBERT dalam klasifikasi dataset *clickbait*. Selain itu, pada skenario terbaik, *back translation* mampu meningkatkan *F1-score* sebesar 7,84% pada dataset *clickbait* dan 8,16% pada dataset *hate speech*.

2.1.4 Penelitian Keempat (Gonzalez-Carvajal dan Garrido-Merchan, 2020)

Gonzalez-Carvajal dan Garrido-Merchan (2020) membandingkan kinerja model BERT dengan metode pembelajaran mesin tradisional dalam klasifikasi teks dengan melakukan serangkaian eksperimen. Terdapat 4 eksperimen yang dilakukan dengan menggunakan dataset yang berbeda. Eksperimen pertama melibatkan analisis sentimen pada dataset *Internet Movie Database* (IMDB) yang berisi kumpulan ulasan film, model-model untuk klasifikasi sentimen pada dataset ini adalah BERT, *Voting Classifier, Logistic Regression, Linear SVC, Multinomial NB, Ridge Classifier,* dan *Passive Aggressive Classifier*. Hasilnya menunjukkan bahwa BERT lebih unggul dibandingkan model lain untuk klasifikasi ulasan film dengan akurasi mencapai 93,87%. Eksperimen kedua yaitu klasifikasi *tweet* dalam dataset RealOrNot yang berisi *tweet* tentang bencana nyata atau bukan bencana, model yang digunakan untuk klasifikasi ini adalah BERT dan H2OAutoML. Hasil menunjukkan BERT mencapai akurasi 83,61%, sedangkan H2OAutoML hanya mencapai akurasi 78,75%. Eksperimen ketiga menggunakan dataset berita dalam bahasa Portugis. Model yang digunakan untuk klasifikasi ini adalah BERT dan Predictor (auto_ml), hasil klasifikasi juga menunjukkan BERT lebih unggul dengan akurasi mencapai 90,93% dibandingkan dengan teknik auto_ml yang hanya mencapai akurasi 84,80%. Eksperimen keempat menggunakan dataset ulasan hotel dalam bahasa Tiongkok. Model yang digunakan untuk klasifikasi ini adalah BERT dan Predictor (auto_ml). Hasilnya juga menunjukkan BERT memperoleh akurasi lebih unggul yaitu sebesar 93,81%, sedangkan metode Predictor (auto_ml) hanya mencapai akurasi 73,99%.

Keunggulan BERT dalam berbagai jenis data dalam eksperimen tersebut yaitu BERT dapat menangkap emosi dan sentimen dalam teks, memahami konteks yang lebih dalam sehingga memungkinkan model untuk belajar dari data yang lebih luas dan beragam, memahami makna dalam berita dengan mempertimbangkan konteks kata dalam kalimat, dan BERT lebih efektif dalam menangkap bahasa yang kompleks, terutama dalam bahasa yang tidak memiliki pemisahan kata, sehingga BERT menjadi pilihan yang lebih baik dibandingkan dengan metode pembelajaran mesin tradisional dalam tugas klasifikasi teks.

2.1.5 Penelitian Kelima (Sirusstara dkk., 2022)

Sirusstara dkk. (2022) menggunakan dataset CLICK-ID yang terdiri dari 15.000 judul berita yang telah dianotasi sebagai *clickbait* dan *non-clickbait*. Data yang digunakan tidak seimbang, oleh karena itu untuk menghindari set data yang tidak seimbang, Sirusstara dkk. (2022) melakukan *undersampling* sehingga menghasilkan 5.297 judul *non-clickbait* dan 3.316 judul *clickbait*. Beberapa model *pre-trained* yang digunakan yaitu IndoBERT, XLM-RoBERTa, BERT, dan Multinomial NB. Hasil menunjukkan bahwa model IndoBERT (indobenchmark/IndoBERT-base-p1) menunjukkan performa terbaik dalam mengklasifikasikan judul *clickbait* dengan akurasi 92%, presisi 92,41%, *recall* 92,23%, dan *F1-score* 92,32%.

2.2 Berita *online*

Media *online* merupakan bentuk laporan peristiwa yang bersifat multimedia, aktualitas, cepat, update, fleksibel, luas, interaktif, dan terdokumentasi yang tersedia di internet. (Romli, 2018). Salah satu bentuk media *online* yaitu situs web berita. Media cetak atau majalah yang memuat edisi *online* disebut situs berita *online*. Informasi dalam berita *online* disajikan melalui *platform digital* seperti situs web dan media sosial. Bentuk berita *online* sangat beragam termasuk artikel, video, gambar, dan infografis. Selain itu, berita *online* mencakup berbagai topik seperti kesehatan, ekonomi, politik, pendidikan, dan lainnya. Berita *online* dapat diakses melalui internet, memungkinkan masyarakat memperoleh informasi secara cepat dan mudah, serta mempengaruhi persepsi publik dalam merespons suatu informasi.

2.3 *Text Mining*

Text mining atau penambangan teks merupakan proses eksplorasi pengetahuan secara mendalam yang berhubungan dengan kumpulan dokumen dari waktu ke waktu menggunakan berbagai alat analisis (Feldman dan Sanger, 2007). *Text mining* melakukan identifikasi pola-pola menarik dalam sumber data dengan tujuan untuk mengekstrak informasi yang berguna. Kumpulan dokumen merupakan sumber data dalam *text mining*, dimana pola-pola relevan terletak pada data teks yang tidak terstruktur dalam dokumen tersebut. Selain itu, *text mining* mengadaptasi berbagai pola yang pertama kali diperkenalkan dalam penelitian penambangan data. Pada *data mining*, data diasumsikan sudah terstruktur, sehingga prapemrosesan hanya berfokus pada pembersihan dan normalisasi data. Sebaliknya, dalam *text mining*, prapemrosesan berfokus pada identifikasi dan ekstraksi fitur-fitur representatif dalam dokumen bahasa alami. Prapemrosesan ini bertujuan mengubah data tidak terstruktur menjadi format yang lebih terstruktur. Teknik yang digunakan dalam text mining memanfaatkan kemajuan dalam pemrosesan bahasa alami, seperti pencarian informasi, ekstraksi informasi, dan linguistik komputasi berbasis korpus.

2.4 *Imbalanced Data*

Ketidakseimbangan data *(imbalanced data)* merupakan masalah yang terjadi karena adanya distribusi kelas yang tidak seimbang dalam data. Kelas yang memiliki jumlah data lebih banyak tergolong dalam mayoritas, sedangkan kelas yang memiliki jumlah data lebih sedikit tergolong dalam minoritas. Apabila terdapat kelas yang mendominasi dataset, maka hal ini menunjukkan ketidakseimbangan dataset yang dapat menyebabkan beberapa permasalahan dalam proses klasifikasi (Hancock dkk., 2024). Ketidakseimbangan kelas dapat menimbulkan bias yang menyebabkan model menjadi condong ke berbagai sisi (Ghosh dkk., 2024). Kondisi ini seringkali mengakibatkan model yang telah dilatih cenderung mengabaikan kelas minoritas. Ketidakseimbangan dataset terjadi dalam berbagai aplikasi, seperti deteksi penipuan, diagnosis medis, dan pengenalan anomali.

2.5 Augmentasi Data

Augmentasi data merupakan teknik dalam pembelajaran mesin terutama pada *Natural Language Processing* (NLP) yang dapat digunakan untuk meningkatkan kumpulan data pelatihan agar model dapat belajar lebih baik dan meningkatkan kinerja model (Jungiewicz dan Smywinski-Pohl, 2019). Selain itu, teknik augmentasi data dapat membantu menambah keragaman data pelatihan yang memungkinkan model digeneralisasikan ke data pengujian yang belum pernah dilihat. Melalui proses ini, dataset yang sebelumnya tidak seimbang dapat diperkuat dengan data tambahan hasil augmentasi, sehingga jumlah data latih menjadi lebih banyak, bervariasi, dan seimbang.

2.5.1 *Synonym Replacement* (SR)

Augmentasi data menggunakan *synonym replacement* merupakan salah satu teknik untuk mengatasi ketidakseimbangan data. Prosesnya berupa penambahan variasi data dengan mengubah n kata menjadi sinonim atau kata yang bermakna sama. Meskipun tidak semua hasil augmentasi dengan sinonim menghasilkan kalimat yang sesuai konteks, metode ini tetap efektif karena memiliki kemampuan yang baik dalam mempertahankan label, mengurangi *overfitting*, dan meningkatkan generalisasi model dengan memperkenalkan variasi dalam data pelatihan (Jungiewicz dan Smywinski-Pohl, 2019). Penggantian kata sinonim didasarkan pada penelitian sebelumnya (Ningsih dkk., 2022) yang dimodifikasi dengan menambahkan POS *tagging* untuk menentukan kata yang akan terpilih sebagai sinonim. Suatu kalimat akan dipecah menjadi token terlebih dahulu kemudian dilakukan pelabelan dengan POS *tagging*.

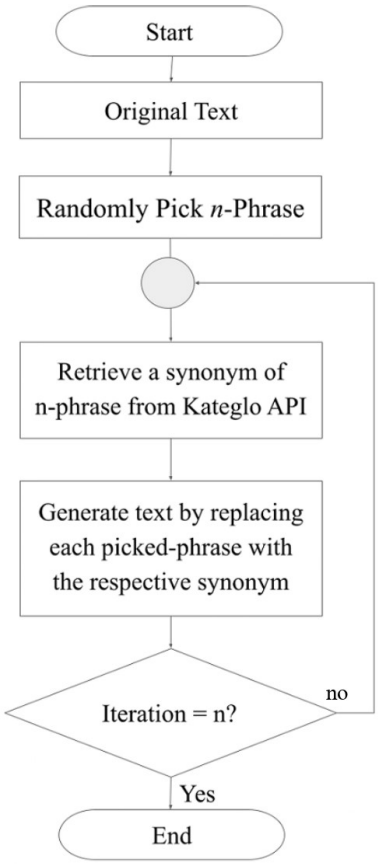
Part-of-Speech (POS) *tagging* merupakan proses dalam NLP yang bertujuan memberikan label kelas kata pada setiap kata dalam suatu kalimat (Zhang dkk., 2021). *Tag* ini dapat menentukan fungsi dari tata bahasa kata tersebut. Salah satu *library* yang dapat melakukan POS *tagging* yaitu *library Stanza*. *Stanza* merupakan sebuah *toolkit open-source* untuk NLP yang mendukung 66 bahasa manusia (Qi dkk., 2020). *Library Stanza* menghasilkan dua tingkatan POS *tag* yaitu:

1. *Universal POS tags* (UPOS) merupakan *tag* universal untuk kelas kata secara umum. Jumlah *tag* UPOS yang umum adalah 17 *tag* seperti ADJ (*adjective*, kata sifat), ADP (*adposition*, preposisi), ADV (*adverb*, kata keterangan),

AUX (*Auxiliary Verb*, kata kerja bantu), CCONJ (*Coordinating Conjunction*, konjungsi koordinatif), DET (*determiner*, kata penentu), INTJ (*interjection*, kata seru), NOUN (kata benda), NUM (*Numeral*, angka), PART (*Particle*, partikel), PRON (*pronoun*, kata ganti), PROPN (*Proper Noun*, Nama Orang, tempat, dan lainnya), PUNCT (*punctuation*, tanda baca), SCONJ *subordinating conjunction*, konjungsi subordinatif), SYM (*symbol*, simbol), VERB (kata kerja), dan X (*other*, kategori lainnya). *Tag* ini telah disesuaikan dengan standar Universal Dependencies (UD).

2. *Language-specific POS tags* (XPOS) merupakan *tag* yang lebih spesifik sesuai bahasanya sehingga jumlah *tag* lebih banyak dan detail dibandingkan UPOS. Jumlah *tag* yang tersedia mencapai 40, beberapa diantaranya yaitu NN (*noun singular*, kata benda tunggal), NNS (*noun plural*, kata benda jamak), VB (*verb base form*, kata kerja bentuk dasar), VBD (*verb past tense*, kata kerja bentuk lampau), JJ (*adjective*, kata sifat), dan *tag* lainnya.

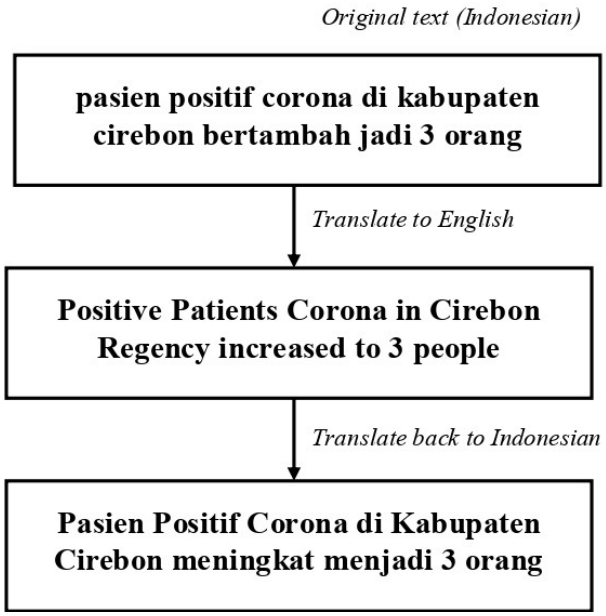
Pada proses augmentasi *synonym replacement*, penggantian kata dilakukan dengan menggunakan daftar kata sinonim yang diperoleh dari API Kateglo yang merupakan layanan web yang menyediakan berbagai informasi linguistik, seperti sinonim, antonim, entri formal, dan glosarium bahasa Indonesia (Rohman dkk., 2019). Kateglo adalah singkatan dari kata ka(mus), te(saurus), dan glo(sarium). Kateglo terdiri dari 72.253 entri kamus, 191.000 entri glosarium, 2.012 entri peribahasa, serta 3.423 entri singkatan dan akronim (Ningsih dkk., 2022). Pencarian sinonim dengan kateglo dapat diakses menggunakan *Application Programming Interface* (API) melalui `kateglo.com/api.php`. Ningsih dkk. (2022) telah mengembangkan metode pembangkitan data yaitu Bangkit-TS (Pembangkit Teks Berbasis Sinonim) untuk mengatasi ketidakseimbangan kelas pada data teks. Metode ini dapat menghasilkan n data baru, prosesnya melibatkan pemilihan n-kata secara acak dalam teks untuk diganti dengan sinonim yang bersumber dari kamus Bahasa Indonesia yang ada di Internet menggunakan Kateglo API. Tahapan pembangkitan data berbasis sinonim dengan kateglo API ditunjukkan Gambar 1.



Gambar 1. Kerangka Kerja Pencarian Sinonim dengan Kateglo API (Ningsih dkk., 2022)

2.5.2 *Back Translation* (BT)

Back translation adalah salah satu metode untuk menghasilkan variasi teks baru dengan menerjemahkan teks asli ke dalam bahasa target lalu diterjemahkan kembali ke bahasa semula (Rahma dan Suadaa, 2023). Proses augmentasi data *back translation* dapat dilakukan melalui pemrograman *Python* dengan memanfaatkan API *Google Translate*. Kalimat yang dihasilkan dari proses ini umumnya akan berbeda secara struktur, namun tetap mempertahankan makna yang sama. Penerapan metode ini cukup efektif dalam bidang NLP terutama untuk meningkatkan performa model dan memperkaya variasi dataset. Mekanisme *back translation* dapat diilustrasikan pada Gambar 2.



Gambar 2. Mekanisme *Back Translation*

2.6 Klasifikasi

Klasifikasi adalah proses menemukan model dengan membedakan kelas data sehingga membentuk kategori atau kelompok yang dapat digunakan untuk memprediksi label kelas yang belum diketahui (Han dan Kamber, 2006). Klasifikasi merupakan bagian dari NLP yang dapat mengelompokan dokumen yang berisi teks yang berbeda (Herrera dkk., 2016). Tujuan klasifikasi adalah untuk mempelajari pola berlabel dari model kemudian memprediksi label untuk sampel data yang belum pernah dilihat. Kumpulan atribut dalam klasifikasi dibagi menjadi dua yaitu fitur *input* yang berisi variabel untuk memprediksi hasil dan label *output* berupa variabel yang diprediksi oleh model. Hubungan antara fitur *input* dan label *output* dapat dianalisis melalui model hasil klasifikasi. Model hasil pelatihan dapat digunakan untuk memproses kumpulan sampel data baru agar mendapatkan prediksi kelas. Dataset biner hanya memiliki satu atribut keluaran dan hanya dapat menerima dua nilai berbeda yaitu positif dan negatif yang dapat diartikan sebagai benar dan salah, 1 dan 0, atau gabungan lain yang terdiri dari 2 nilai. Beberapa penerapan klasifikasi biner diantaranya klasifikasi *spam* email, analisis pinjaman, kredibilitas pelanggan, evaluasi medis, dan pengenalan berbagai pola biner lainnya.

2.7 Fungsi Aktivasi

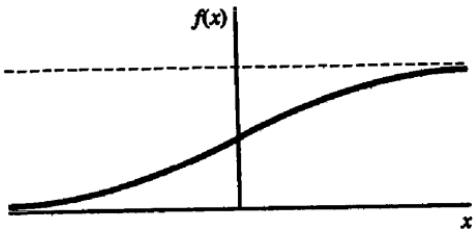
Fungsi aktivasi pada jaringan saraf digunakan untuk menghitung jumlah bobot dari *input* dan bias. Fungsi ini mengolah data melalui penurunan gradien sehingga menghasilkan *output* yang mencakup parameter dari data tersebut. Fungsi aktivasi juga disebut fungsi *transfer*. Fungsi aktivasi dapat mengatur *output* jaringan saraf di berbagai bidang, seperti pengenalan dan klasifikasi objek, pengenalan suara, perkiraan cuaca, deteksi kanker, dan terjemahan mesin (Nwankpa dkk., 2018).

2.7.1 Fungsi Aktivasi Sigmoid

Fungsi aktivasi sigmoid memiliki kurva berbentuk S dan bersifat *non-linear* yang memiliki *range* dari $(-\infty, +\infty)$ ke $[0, 1]$ (Rasamoelina dkk., 2020). Fungsi aktivasi sigmoid yang digunakan dalam klasifikasi biner juga dapat menghasilkan nilai *output* biner atau berada dalam interval antara 0 dan 1. Fungsi ini disebut sebagai sigmoid biner (Fauset, 1994). Fungsi aktivasi sigmoid dapat didefinisikan dalam persamaan (2.1) dan diilustrasikan oleh Gambar 3.

$$f(x) = (\frac{1}{1 + exp^{-x}}) \tag{2.1}$$

Keterangan:
x = nilai *input*
exp = nilai eksponensial (2.7183)
f(x) = *output* dari fungsi Sigmoid



Gambar 3. Fungsi Aktivasi Sigmoid (Fauset, 1994)

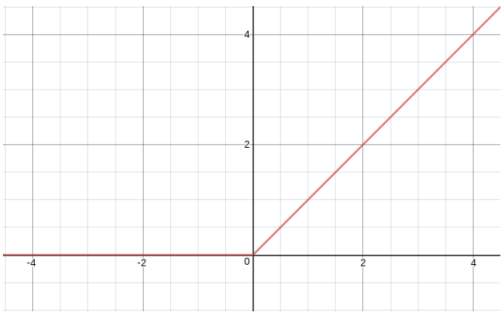
2.7.2 Fungsi Aktivasi ReLU

Fungsi Aktivasi *Rectified Linear Unit* (ReLU) merupakan fungsi aktivasi yang lebih cepat dan mudah dioptimalkan dengan metode gradien turunan karena mempertahankan sifat-sifat model linier sehingga memberikan generalisasi yang lebih baik dalam pembelajaran mendalam (Nwankpa dkk., 2018). Fungsi aktivasi ReLU didefinisikan oleh persamaan (2.2) sebagai berikut.

$$f(x) = \max(0, x) = \begin{cases} x_i, & if\ \mathrm{x}_i \geq 0 \\ 0, & if\ \mathrm{x}_i < 0 \end{cases} \tag{2.2}$$

Keterangan:
x = Nilai *input*
f(x) = *Output* dari fungsi ReLU

Fungsi aktivasi ReLU mengatur agar setiap elemen *input* yang kurang dari nol diubah menjadi nol. Fungsi ini merupakan fungsi linear untuk semua nilai positif dan nol untuk semua nilai negatif. Selain itu, fungsi aktivasi ReLU dapat mengaktifkan lapisan tersembunyi dalam jaringan saraf yang mengandung satu atau lebih nilai nol yang sebenarnya (Rasamoelina dkk., 2020). Fungsi aktivasi ReLU diilustrasikan oleh Gambar 4.



Gambar 4. Fungsi Aktivasi ReLU (Agarap, 2018)

2.8 *Fine-Tuning*

Fine-tuning adalah proses melatih ulang model yang sudah ada pada dataset baru. *Hyperparameter* merupakan nilai parameter yang tidak berubah dan telah ditetapkan sebelum pelatihan model. Pengaturan proses pelatihan dan kekuatan model membuat *hyperparameter* mempengaruhi kinerja model (Xu dkk., 2023).

Proses pengoptimalan *hyperparameter* disebut *hyperparameter tuning* yang menentukan nilai *hyperparameter* dengan kinerja prediksi terbaik dari semua nilai *hyperparameter* yang diuji. *Hyperparameter* dilakukan dengan menambahkan beberapa parameter seperti *batch size*, *epoch*, *dropout*, *learning rate*, *Adam optimizer*, dan *loss function*. Devlin dkk. (2019) menyatakan bahwa terdapat nilai parameter yang harus disesuaikan nilainya agar menghasilkan kinerja yang optimal dalam *fine-tuning* seperti *batch size*, *learning rate*, dan jumlah *epoch*. Berikut beberapa *hyperparameter* yang digunakan dalam model:

a) *Batch size* (ukuran *batch*) adalah jumlah sampel yang digunakan dalam satu iterasi selama pelatihan model sebelum pembaruan bobot (Xu dkk., 2023). Ukuran *batch* mempengaruhi kecepatan pelatihan dan stabilitas pembaruan.
b) *Epoch* merupakan jumlah yang mengatur berapa kali model belajar dari data. Jumlah *epoch* yang terlalu rendah dapat menyebabkan *underfitting*, sedangkan jumlah *epoch* yang terlalu tinggi dapat menyebabkan *overfitting* dan memperlambat waktu komputasi dengan hasil yang lebih buruk (Xu dkk., 2023).
c) *Dropout* merupakan teknik untuk mencegah *overfitting* dengan menonaktifkan sejumlah *neuron* selama pelatihan berdasarkan nilai yang telah ditetapkan, sehingga mengurangi ketergantungan antar *neuron* (Xu dkk., 2023).
d) *Learning rate* merupakan *hyperparameter* untuk mengatur besar perubahan dalam pembaruan bobot selama proses optimisasi (Xu dkk., 2023). *Learning rate* yang terlalu tinggi dapat menyebabkan model tidak stabil dan gagal mencapai konvergensi yang diinginkan, sementara *learning rate* yang terlalu rendah dapat memperlambat proses pelatihan. Oleh sebab itu, penting untuk menentukan nilai *learning rate* yang tepat agar menghasilkan proses pelatihan yang optimal.
e) *Adaptive Moment Estimation* (Adam) *optimizer* adalah metode untuk optimisasi stokastik yang memerlukan gradien orde pertama dengan memori yang sedikit. *Optimizer Adam* menghitung estimasi adaptif untuk setiap parameter berdasarkan momen pertama dan kedua dari gradien (Kingma dan Ba, 2015). Metode optimasi *Adam* juga bersifat serbaguna dan efisien sehingga membuatnya berfungsi dengan baik dalam berbagai model terutama dalam banyak aplikasi deep learning seperti mengukur performa suatu model dalam memprediksi label.
f) Fungsi yang menggabungkan *Binary Cross Entropy* (BCE) dan fungsi sigmoid ke dalam satu komputasi tunggal yang lebih stabil secara numerik daripada menggunakan sigmoid biasa yang diikuti oleh *BCELos*s dikenal dengan fungsi *BCEWithLogitsLoss* (Babaali dkk., 2022). Pada klasifikasi biner atau prediksi

dua kelas, fungsi *BCEWithLogitsLoss* dapat digunakan sebagai fungsi kerugian untuk menentukan perbedaan antara hasil prediksi dan label yang sebenarnya. *Output* model yang dihasilkan fungsi ini berupa nilai logit dapat langsung digunakan tanpa diubah menjadi probabilitas. Fungsi *BCEWithLogitsLoss* didefinisikan oleh persamaan (2.3) dan persamaan (2.4).

$$BCE = -\sum_{i=1}^{C=2} t_i \log(f(s_i)) \tag{2.3}$$

$$BCEwithLogitsLoss = -t_1 \log(f(s_1)) - t_2 \log(f(s_2)) \tag{2.4}$$

Keterangan:
s_i adalah nilai logit untuk kelas C_i
$f(s_i)$ = $f(x)$ adalah *output* fungsi sigmoid pada persamaan (2.1)
t_i adalah nilai kebenaran *(ground truth)* untuk kelas ke-i
i adalah jumlah kelas
Apabila terdapat dua kelas C_1 dan kelas C_2, maka:
$t_1 \in \{0, 1\}$ adalah nilai kebenaran untuk kelas C_1
s_1 adalah skor untuk kelas C_1
$t_2 = 1 - t_1$ adalah nilai kebenaran untuk kelas C_2
$\mathrm{s}_2 = 1 - s_1$ adalah skor untuk kelas C_2

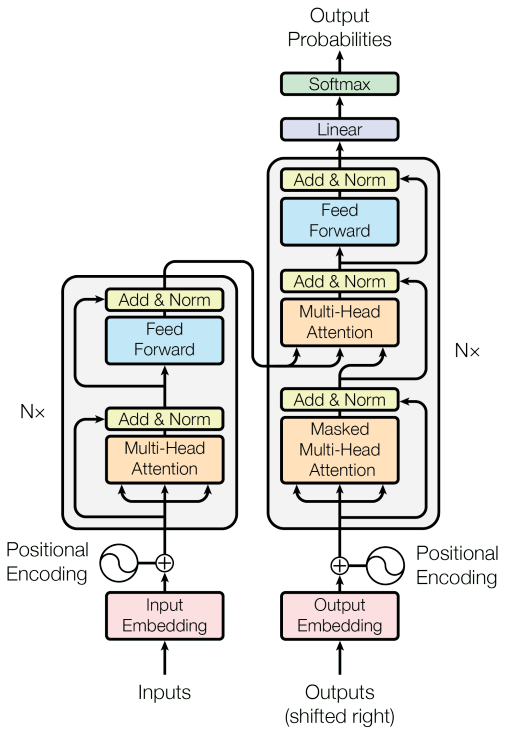
2.9 *Transformer*

Recurrent Neural Networks (RNN) atau jaringan saraf berulang merupakan jenis jaringan saraf tiruan yang digunakan untuk memproses data urutan seperti teks. Namun, RNN memiliki keterbatasan dalam memproses urutan kata sehingga menghalangi pelatihan secara paralel, hal ini disebabkan oleh sifat sekuensial RNN, dimana setiap posisi kata dalam urutan bergantung pada hasil sebelumnya. Arsitektur jaringan baru yang didasarkan pada mekanisme perhatian dan tidak menggunakan lapisan *recurrent* atau konvolusi dikenal dengan *transformer*. Model ini telah dilatih selama dua belas jam pada delapan *Graphics Processing Unit NVIDIA Tesla P100* (GPU P100) sehingga memungkinkan paralelisasi yang jauh lebih banyak pada saat pelatihan dibandingkan model *recurrent* yang sering kali terhambat oleh sifat sekuensial dari proses pelatihannya. *Self-Attention* dalam *transformer* dapat menghubungkan posisi berbeda dari satu urutan untuk

menghasilkan representasi dari urutan tersebut. Mekanisme ini telah berhasil diterapkan dalam berbagai tugas, termasuk pemahaman bacaan, ringkasan abstrak, *textual entailment* dan representasi kalimat (Vaswani dkk., 2017).

2.9.1 *Encoder* dan *Decoder*

Transformer merupakan model transduksi pertama yang sepenuhnya mengandalkan *self-attention* untuk menghitung representasi *input* dan *output* dengan menggunakan struktur *encoder-decoder*. *Encoder* berfungsi mengonversi urutan *input* simbol $x_1, \ldots, x_n$ menjadi representasi kontinu $z = (z_1, \ldots, z_n)$. *Decoder* memproses simbol secara bertahap dengan merepresentasikan z untuk menghasilkan *output* $y_1, \ldots, y_n$. *Encoder* dan *decoder* tersusun dari 6 lapisan identik. Setiap lapisan *encoder* memiliki 2 sub-lapisan yaitu *multi-head self-attention* dan *feed forward network*. Sedangkan *decoder* memiliki 3 sub-lapisan yaitu *masked multi-head self-attention, multi-head self-attention* dan *feed forward network*. Setiap sub-lapisan *encoder* dan *decoder* dilengkapi dengan koneksi residual dan normalisasi lapisan (Vaswani dkk., 2017). Arsitektur model *transformer* ditunjukkan pada Gambar 5. sebagai berikut:



Gambar 5. Arsitektur *Transformer* (Vaswani dkk., 2017)

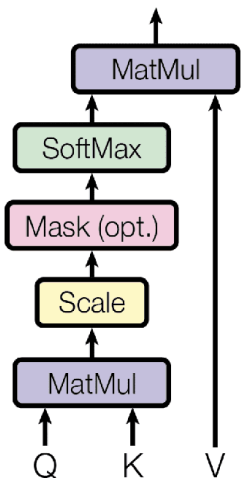
2.9.2 *Attention*

Fungsi *Attention* merupakan suatu mekanisme yang menghubungkan sebuah *query* dan sekumpulan pasangan *key-value* ke dalam *output*. Mekanisme dilakukan dengan menghitung *output* sebagai jumlah nilai bobot yang ditentukan oleh kesesuaian antara *query* dan kunci. Hal ini memungkinkan *query* untuk fokus pada bagian tertentu dari *input* sehingga dapat meningkatkan efisiensi dalam menangkap informasi relevan dari konteks yang lebih luas (Vaswani dkk., 2017).

2.9.3 *Scaled Dot-Product Attention*

Scaled Dot-Product Attention memiliki vektor *query* dan *key* yang berdimensi d_k, serta vektor value yang berdimensi d_v. Prosesnya melibatkan perkalian antara *query* dan *key*, kemudian dibagi dengan akar kuadrat dari dimensi *key* untuk menghindari nilai yang terlalu besar dan menerapkan fungsi *softmax* untuk menghasilkan bobot pada *value*. *Query* (Q) adalah vektor yang digunakan untuk mencari relevansi dalam urutan *input* dan *Key* (K) adalah kunci untuk menyesuaikan relevansi, dan *Value* (V) adalah nilai atau informasi yang digunakan untuk menghasilkan *output* (Vaswani dkk., 2017). Fungsi *attention* dapat dihitung dengan menggunakan persamaan (2.5). Selain itu *scaled dot-product attention* ditunjukkan oleh Gambar 6. sebagai berikut:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.5}$$



Gambar 6. *Scaled Dot-Product Attention* (Vaswani dkk., 2017)

2.9.4 *Multi-Head Attention*

Multi-head Attention adalah pendekatan yang memperluas *attention* dengan melakukan beberapa fungsi *attention* secara paralel yang menghasilkan dimensi d_v. Mekanisme ini memproyeksikan *query, key,* dan *value* ke dalam dimensi yang berbeda untuk setiap ”h”. *Multi-head attention* berguna untuk mengumpulkan informasi dari berbagai representasi secara bersamaan (Vaswani dkk., 2017). Perhitungan *multi-head attention* didefinisikan pada persamaan(2.6) dan diilustrasikan pada Gambar 7.

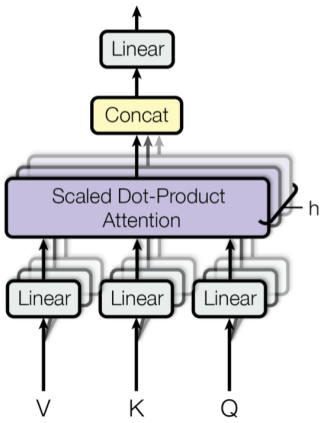
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.6}$$

dengan:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Proyeksi didefinisikan oleh matriks parameter

$$W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}, \quad W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$$



Gambar 7. *Multi-Head Attention* (Vaswani dkk., 2017)

2.9.5 *Position-wise Feed-Forward Networks*

Feed-Forward Network (FFN) merupakan bagian yang terdapat dalam setiap lapisan *encoder* dan *decoder* pada arsitektur *transformer*. Proses dalam FFN bersifat terhubung penuh dan dapat diterapkan secara terpisah dan identik. FFN dihubungan oleh fungsi aktivasi ReLU dengan 2 transformasi linear yang sama. Meskipun demikian, terdapat perbedaan parameter yang digunakan dalam setiap

lapisannya (Vaswani dkk., 2017). FFN didefinisikan oleh persamaan (2.7).

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2.7}$$

2.9.6 *Positional Encoding*

Positional encoding perlu ditambahkan ke dalam *embedding input* bagian dasar *encoder* dan *decoder* model karena *transformer* tidak menggunakan *recurrent* dan konvolusi. Melalui *positional encoding*, model dapat menggunakan urutan kata dengan memasukkan informasi posisi token dalam kata. Dimensi dalam *positional encoding* sama dengan *embedding* sehingga keduanya dapat dijumlahkan (Vaswani dkk., 2017). *Positional encoding* didefinisikan oleh persamaan (2.8) dan persamaan (2.9) sebagai berikut:

$$\text{PE}_{\text{pos},2i} = \sin\left(\text{pos}/10000^{\frac{2i}{d_{\text{model}}}}\right) \tag{2.8}$$

$$\text{PE}_{\text{pos},2i+1} = \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{2.9}$$

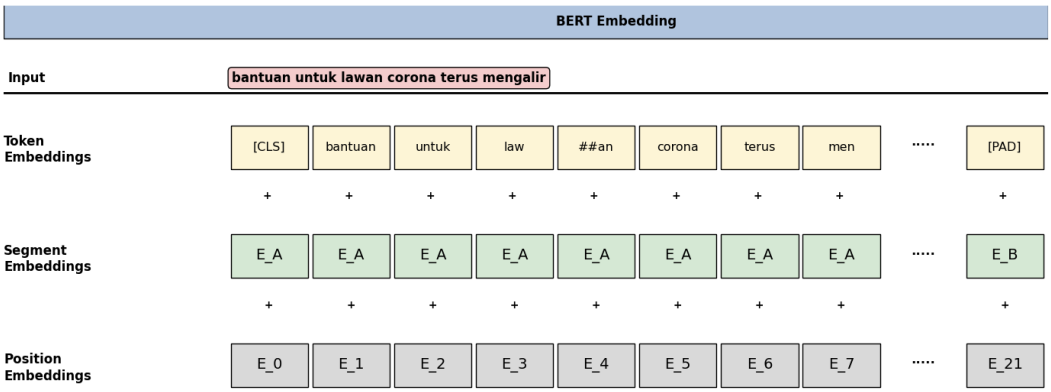
2.10 *Pre-Trained Transformer Models*

Pada model yang menggabungkan mekanisme perhatian antara *encoder* dan *decoder*, *transformer* dengan cepat menggantikan LSTM dan CNN. *Pre-trained transformer models* merupakan pendekatan baru dalam pemrosesan bahasa alami yang memanfaatkan arsitektur *transformer* berdasarkan *attention mechanism* yang membuat model yang telah dilatih sebelumnya dapat menghasilkan sematan kata untuk memahami dan mengklasifikasikan teks. Arsitektur model *transformer* umumnya mengikuti struktur dasar pada BERT terutama dalam mekanisme embedding model. Perbedaan utama diantara berbagai model *pre-trained* lain terletak pada cara tokenisasi yang disesuaikan dengan karakteristik korpus masing-masing model. Arsitektur BERT dan model-model *transformer* sejenis memiliki token *classification* [CLS] yang digunakan sebagai awalan dalam proses *embedding*. Kemudian token *separator* [SEP] untuk memisahkan kalimat. Token *padding* [PAD] untuk menyamakan panjang *input* dengan mengisi token yang kosong. Tokenisasi yang dilakukan menggunakan metode *word piece tokenization*

yang memanfaatkan kosakata dalam model serta kosakata tambahan diluar model (Devlin dkk., 2019). Sebelum ke tahap ekstraksi fitur, dilakukan 2 tahap *embedding* yaitu *segment embedding* dan *positional embedding*. Pada *segment embedding*, jika *input* hanya terdiri dari satu kalimat dan representasi vektor hanya berupa indeks 0 dan indeks 1 maka *segment embedding* yang digunakan adalah indeks 0. Sedangkan *positional embedding* menggunakan tabel pencarian berukuran (n, 768), dengan n adalah panjang kalimat. Baris pertama mewakili representasi vektor untuk kata pada posisi pertama, baris kedua untuk kata pada posisi kedua, dan seterusnya. Gabungan dari ketiga proses *embedding* ini disebut *input embedding*.

2.10.1 *Bidirectional Encoder Representations from Transformers* (BERT)

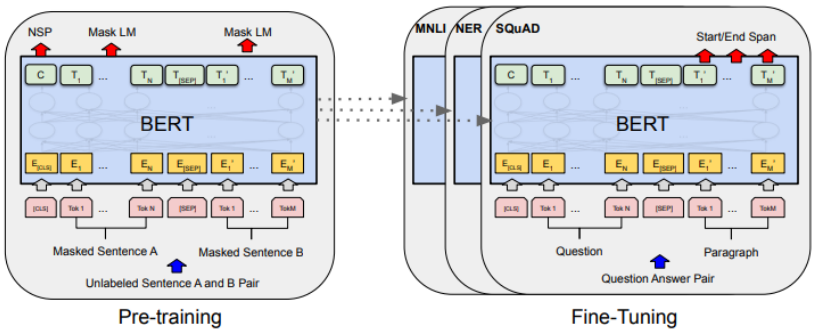
Bidirectional Encoder Representations from Transformers (BERT) merupakan *transformer* dua arah yang dilatih berdasarkan pemodelan bahasa pada teks yang memiliki korpus besar. Melalui penerapan pelatihan dua arah, model bahasa yang dilatih dapat memiliki pemahaman yang lebih mendalam (Devlin dkk., 2019). Model BERT telah dilatih pada 2,5 miliar kata dari Wikipedia bahasa Inggris, dan 800 juta korpus buku. *BERT Base Multilingual Cased* adalah salah satu versi model BERT yang tersedia dalam *hugging face* yang memiliki 12 lapisan *encoder*, 768 dimensi tersembunyi, 110 juta parameter, 768 ukuran *embedding*, 110.000 juta parameter, dan ukuran kosakata 30.000 token serta memiliki tipe model *case-sensitive* artinya dapat membedakan huruf besar dan kecil. Representasi BERT *embedding* diilustrasikan oleh Gambar 8.



Gambar 8. Arsitektur BERT *Embedding*

Pre-training model BERT merupakan proses melatih model untuk memahami

konteks kata dalam kalimat dengan cara mendalam dan *bidirectional* (Devlin dkk., 2019). Proses *pre-training* BERT melibatkan *Masked Language Model* (MLM) dan *Next Sentence Prediction* (NSP). Model ini dilatih pada korpus teks besar seperti Wikipedia dan korpus buku. Kemampuan BERT dalam menangkap hubungan kompleks antar kata dan kalimat sehingga sangat efektif untuk berbagai tugas pemrosesan bahasa alami berasal dari arsitektur *transformer* yang digunakannya. Setelah tahap *pre-training*, model yang telah dilatih kemudian disesuaikan *(fine-tuning)* untuk tugas tertentu menggunakan dataset berlabel. Pada proses ini, model dioptimalkan untuk menyelesaikan tugas tertentu, seperti klasifikasi teks, pengenalan entitas, atau pertanyaan dan jawaban. Proses *fine-tuning* melibatkan penyesuaian parameter berdasarkan data yang ada hingga model dapat memberikan hasil yang lebih akurat dan optimal. Proses ini lebih cepat dan memerlukan lebih sedikit data dibandingkan dengan pelatihan dari awal. Ilustrasi proses *pre-training* dan *fine-tuning* dapat dilihat pada Gambar 9.



Gambar 9. *Pre-training* dan *Fine-tuning* (Devlin dkk., 2019)

2.10.1.1. *Masked Language Model* (MLM)

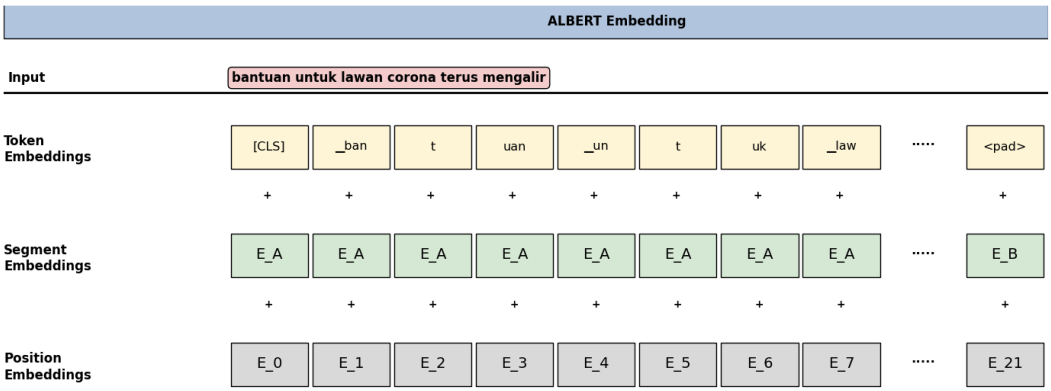
Masked Language Model adalah tugas utama dalam *pre-training* BERT yang memungkinkan model untuk belajar representasi kata berdasarkan konteks di sekitarnya (Devlin dkk., 2019). Terdapat beberapa kata dalam kalimat secara acak yang disembunyikan *(masked)* dalam MLM, kemudian model akan dilatih untuk memprediksi kata-kata yang hilang tersebut berdasarkan kata-kata yang tersisa. Melalui pendekatan ini, BERT dapat memanfaatkan informasi dari kedua sisi (kiri dan kanan) dari kata yang disembunyikan, yang berbeda dari model bahasa tradisional yang hanya memproses teks secara sekuensial dari kiri ke kanan. Pendekatan ini memungkinkan BERT untuk menghasilkan representasi yang lebih kaya dan lebih akurat dari kata-kata dalam konteks yang lebih luas.

2.10.1.2. *Next Sentence Prediction* (NSP)

Next Sentence Prediction merupakan proses dalam pelatihan model BERT yang bertujuan untuk memahami hubungan antara dua kalimat (Devlin dkk., 2019). Penggunaan NSP memungkinkan model harus memprediksi sepasang kalimat, apakah kalimat kedua mengikuti kalimat pertama dalam konteks yang sama. Setiap pasangan kalimat memiliki 50% kemungkinan menjadi pasangan yang benar dan 50% kemungkinan menjadi pasangan acak.

2.10.2 *A Lite BERT* (ALBERT)

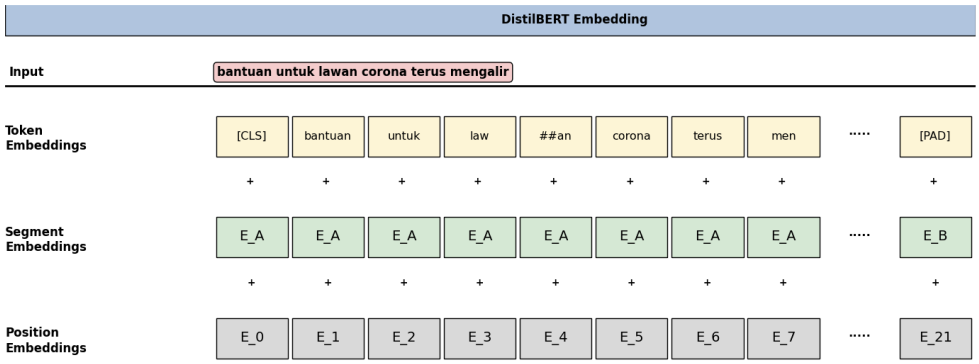
Model ALBERT merupakan pengembangan model BERT yang lebih sederhana dengan menggunakan teknik pengurangan dua parameter untuk meningkatkan kecepatan pelatihan dan konsumsi memori (Gupta dkk., 2021). Struktur utama yang membedakan ALBERT dari BERT adalah penggunaan *factorized embedding parameterization* yang memisahkan ukuran kosakata *embedding* dari ukuran lapisan tersembunyi untuk mengurangi jumlah parameter, kemudian ALBERT juga menggunakan parameter sharing di seluruh lapisan *(cross-layer parameter sharing)* yang memungkinkan parameter dibagi di seluruh lapisan sehingga model menjadi lebih kecil dan efisien. Model ALBERT tidak menggunakan NSP, namun menggantinya dengan *Sentence Order Prediction* (SOP) yang dirancang untuk menangkap koherensi antar kalimat, bukan sekadar prediksi topik seperti pada NSP. Pendekatan SOP bekerja dengan menggunakan dua segmen kalimat berturut-turut dari dokumen yang sama sebagai contoh positif, sementara contoh negatif dihasilkan dengan menukar urutan kedua segmen tersebut. Pendekatan ini mendorong model untuk memahami hubungan logis dan struktur wacana secara lebih mendalam. Model *ALBERT Base* merupakan versi ALBERT dengan jumlah parameter lebih sedikit namun lebih efisien dalam hal komputasi dan memori, model ini menggunakan 12 lapisan *encoder*, 768 dimensi tersembunyi, 128 ukuran *embedding*, dan 12 juta parameter (Lan dkk., 2019). Selain itu, model tersedia dalam *hugging face* dan telah dilatih dengan data korpus buku dan Wikipedia dalam berbagai bahasa, termasuk Bahasa Indonesia (Nabiilah dan Suhartono, 2023). Representasi ALBERT *embedding* diilustrasikan oleh Gambar 10.



Gambar 10. Arsitektur ALBERT *Embedding*

2.10.3 *Distilled BERT* (DistilBERT)

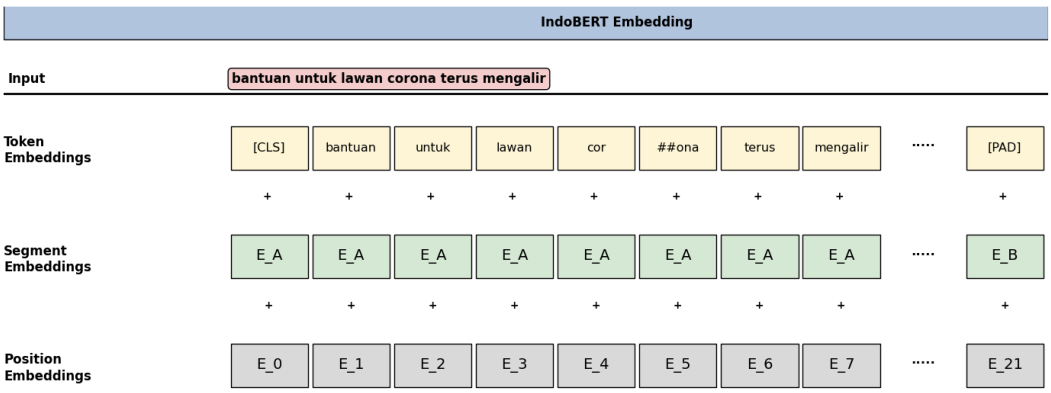
Model DistilBERT merupakan versi BERT yang lebih ringan, cepat, dan menggunakan teknik distilasi selama proses pelatihan sehingga dapat mengurangi ukuran model sebesar 40% dan kecepatan 60% lebih tinggi serta pemahaman kemampuan bahasa hingga 97% (Sanh dkk., 2019). Model DistilBERT mempertahankan 50% lapisan, menghapus *pooler* dan token embedding dari arsitektur BERT. Model ini dilatih menggunakan kombinasi tiga jenis loss yaitu *language modeling loss*, *distillation loss*, dan *cosine-distance loss*. Pendekatan ini memungkinkan model yang lebih kecil tetap mampu melakukan berbagai tugas NLP secara efektif. *DistilBERT Base* adalah salah satu versi DistilBERT yang tersedia dalam *hugging face* yang memiliki 6 lapisan *encoder*, 768 dimensi tersembunyi, 768 ukuran embedding, 66 juta parameter, 104 Bahasa (termasuk Bahasa Indonesia), dan memiliki tipe model *case-sensitive*. Ukuran model ini lebih kecil dibandingkan BERT sehingga lebih cepat dan hemat memori. Representasi DistilBERT *embedding* diilustrasikan oleh Gambar 11.



Gambar 11. Arsitektur DistilBERT *Embedding*

2.10.4 *Indonesian BERT* (IndoBERT)

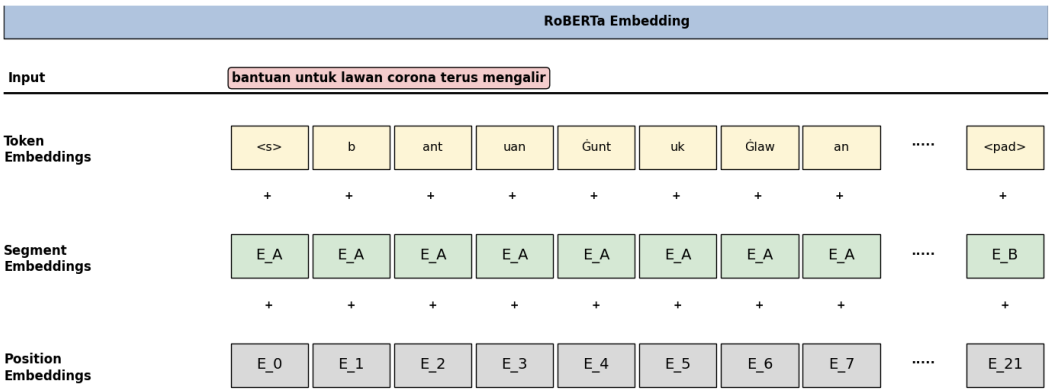
Model IndoBERT adalah model berbasis BERT yang dilatih menggunakan *masked language modeling* khusus untuk tugas bahasa Indonesia yang terdiri lebih dari 220 juta kata dengan tiga sumber utama yaitu Wikipedia (74 juta), artikel berita dari Kompas, Titto, dan Liputan6 (55 juta), serta web korpus (90 juta) (Nabiilah dkk., 2023). *IndoBERT Base* adalah salah satu versi IndoBERT yang tersedia dalam *hugging face* yang memiliki 12 lapisan *encoder*, 768 dimensi tersembunyi, 768 ukuran *embedding*, dan 124,5 juta parameter (Wilie dkk., 2020). Representasi IndoBERT *embedding* diilustrasikan oleh Gambar 12.



Gambar 12. Arsitektur IndoBERT *Embedding*

2.10.5 *Robustly Optimized BERT Pretraining Approach* (RoBERTa)

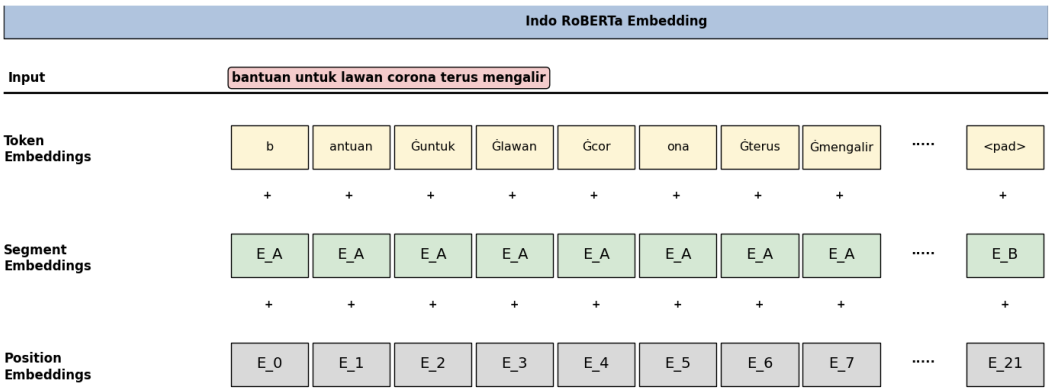
Model RoBERTa merupakan pengembangan model terlatih berbasis *transformer* oleh tim Facebook AI yang menggunakan lebih banyak data dengan ukuran yang lebih kecil dan waktu yg lebih lama (Gupta dkk., 2021). Model RoBERTa telah dilatih pada 160 *gigabytes* korpus dataset *Common Crawl Nes* dan Wikipedia Bahasa Inggris. Selain itu, model RoBERTa mampu mengatasi masalah NSP yang ada di BERT dan menggunakan *masking* dinamis selama pelatihan sehingga set token yang di *mask* akan berubah dan berbeda di setiap epoch pelatihan. *Byte-Pair Encoding* (BPE) adalah jenis tokenisasi yang memiliki ukuran kosakata 50.000 token dan digunakan oleh RoBERTa. *RoBERTa Base* merupakan salah satu versi RoBERTa yang tersedia dalam *hugging face* yang memiliki 12 lapisan *encoder*, 768 dimensi tersembunyi, serta 125 juta parameter (Liu dkk., 2019). Representasi RoBERTa *embedding* diilustrasikan oleh Gambar 13.



Gambar 13. Arsitektur RoBERTa *Embedding*

2.10.6 *Indonesian RoBERTa* (Indo RoBERTa)

Model Indo RoBERTa merupakan pengembangan RoBERTa yang dilatih khusus untuk Bahasa Indonesia dan arsitekturnya terdiri dari 12 lapisan *encoder*, setiap *encoder* memiliki struktur yang terdiri dari 8 lapisan dengan 4 lapisan *multi-head self-attention* dan 4 tingkat *feed-forward* (Widarmanti dkk., 2022). Model RoBERTa yang dilatih dalam 166 bahasa termasuk Bahasa Indonesia pada dataset *Open Super-Large Open Super-Large Crawled ALManaCH corpus* (OSCAR) disebut Indo RoBERTa Base (Nabiilah dan Suhartono, 2023). Representasi Indo RoBERTa *embedding* diilustrasikan oleh Gambar 14.



Gambar 14. Arsitektur Indo RoBERTa *Embedding*

2.11 Evaluasi Model

*P*roses evaluasi model bertujuan menentukan kinerja dari model yang dihasilkan. *Confusion matrix* merupakan salah satu metode evaluasi berupa tabel yang menyajikan jumlah prediksi benar dan salah yang dilakukan oleh model untuk setiap kelas. Tabel 2. menunjukkan *confusion matrix* untuk klasifikasi 2 kelas. *True Negative* (TN) menyatakan jumlah data dalam kelas negatif yang diklasifikasikan dengan benar, *False Positif* (FP) menyatakan jumlah data yang sebenarnya negatif namun salah diklasifikasikan sebagai positif, *False Negatif* (FN) menyatakan jumlah data yang sebenarnya positif namun salah diklasifikasikan sebagai negatif, dan *True Positive* (TP) menyatakan jumlah data dalam kelas positif yang diklasifikasikan dengan benar (Wardhani dkk., 2019). Beberapa metrik evaluasi yang digunakan adalah akurasi, presisi, *recall*, dan *F1-score*. Terdapat 4 kemungkinan yang ditampilkan pada Tabel 2. berikut:

Tabel 2. *Confusion Matrix*

Confusion Matrix	*Predicted Negative*	*Predicted Positive*
Actual Negative	TN	FP
Actual Positive	FN	TP

Beberapa metrik evaluasi dapat dihitung melalui *confusion matrix* seperti akurasi, presisi, *recall*, dan *F1-score*. Namun karena dataset yang digunakan tidak seimbang maka nilai ROC-AUC akan ditambahkan ke dalam metrik evaluasi untuk memberikan gambaran kinerja model yang lebih menyeluruh. Beberapa metrik evaluasi yang digunakan diantaranya sebagai berikut:

1. *Accuracy*

Akurasi menunjukkan proporsi data yang diklasifikasikan dengan benar, baik ke dalam kelas positif ataupun negatif (Nabiilah dan Suhartono, 2023). Rumus akurasi dapat dilihat pada persamaan (2.10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

2. *Precision*

Presisi adalah tingkat keakuratan model dalam melabeli data positif dari total data positif (Nabiilah dan Suhartono, 2023). Rumus presisi dapat dilihat pada persamaan (2.11).

$$Precision = \frac{TP}{TP + FP} \tag{2.11}$$

3. *Recall*

Recall disebut juga sebagai *sensitivity* adalah tingkat keberhasilan model dalam memberikan label yang benar pada data positif (Nabiilah dan Suhartono, 2023). Rumus *recall* dapat dilihat pada persamaan (2.12).

$$recall = \frac{TP}{TP + FN} \tag{2.12}$$

4. *F1-score*

F1-score adalah ukuran kinerja yang mengkombinasikan *precision* dan nilai *recall* (Nabiilah dan Suhartono, 2023). Rumus *F1-score* dapat dilihat pada persamaan (2.13).

$$F1\text{-}score = \frac{2 \times Precision \times recall}{Precision + recall} \tag{2.13}$$

5. *Receiver Operating Characteristics-Area Under Curve (ROC-AUC)*

Receiver Operating Characteristics (ROC) *curve* adalah sebuah grafik yang menggambarkan hubungan antara sensitivitas *(true positive rate)* dan 1 - spesifisitas *(false positive rate)* pada berbagai nilai ambang *(cut off)* dari suatu pengujian. Kurva ROC membantu memilih nilai terbaik untuk mengidentifikasi suatu kondisi. Sementara *Area Under Curve* (AUC) adalah luas area dibawah kurva ROC yang dapat mengukur kemampuan keseluruhan dari suatu pengujian (Bekkar dkk., 2013). Nilai AUC berkisar 0 sampai 1, dimana nilai 1 menunjukkan bahwa pengujian tersebut memiliki kemampuan identifikasi sangat baik. Kategori nilai AUC dapat dilihat pada Tabel 3. sebagai berikut.

Tabel 3. Klasifikasi Nilai AUC

Performance (AUC)	**Klasifikasi**
0.90 – 1.00	Sangat Baik
0.80 – 0.90	Baik
0.70 – 0.80	Rata-rata
0.60 – 0.70	Rendah
0.50 – 0.60	Gagal

Perhitungan AUC yang paling banyak digunakan yaitu dengan *trapezoidal method*. Pendekatan ini menggunakan metode geometris berdasarkan interpolasi linier antara masing-masing titik pada kurva ROC. Pada kasus biner, nilai AUC dapat dihitung dengan *balanced accuracy* pada persamaan (2.14) sebagai berikut (Bekkar dkk., 2013).

$$AUC = \frac{sensitivity + specificity}{2} \tag{2.14}$$

Nilai *sensitivity* dapat mengukur kemampuan model mengenali kelas positif yang dapat dihitung dengan persamaan (2.12), sedangkan *specificity* adalah tingkat kemampuan model mengenali kelas negatif dengan benar yang dapat dihitung dengan persamaan (2.15) sebagai berikut:

$$Specificity = \frac{TN}{TN + FP} \tag{2.15}$$

BAB III

METODE PENELITIAN

3.1 Waktu dan Tempat Penelitian

3.1.1 Tempat Penelitian

Penelitian ini dilaksanakan pada semester genap tahun ajaran 2023/2024 di Pusat Riset Sains Data dan Informasi, Badan Riset dan Inovasi Nasional (BRIN) yang beralamatkan di Jl. Sangkuriang, Dago, Kecamatan Coblong, Kota Bandung, Jawa Barat. Kemudian, penelitian dilanjutkan pada tahun ajaran 2024/2025 di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang beralamatkan di Jalan Prof. Dr. Ir. Soemantri Brojonegoro, Gedong Meneng, Kecamatan Rajabasa, Kota Bandar Lampung, Lampung.

3.1.2 Waktu Penelitian

Kegiatan penelitian dimulai pada semester genap tahun ajaran 2023/2024 tepatnya pada bulan Juni tahun 2024. Tahapan dalam penelitian dibagi menjadi 3 bagian utama. Pertama menentukan topik skripsi, mengumpulkan bahan materi secara studi literatur, mempersiapkan data, dan mengeksplorasi data, kemudian menyusun *draft* proposal penelitian. Tahap kedua mencakup proses pengolahan data dan program yang dimulai dengan *input data, preprocessing data, splitting data, data augmentation, tokenization, text representations,* membangun *pre-trained transformer models*, *fine-tuning models*, dan evaluasi kinerja model berdasarkan metrik evaluasi yang digunakan kemudian membandingkan hasilnya. Pada tahap ketiga dilakukan penyusunan hasil penelitian dan membuat keputusan dari hasil penelitian.

3.2 Data dan Alat Penelitian

3.2.1 Data Penelitian

Data yang digunakan pada penelitian ini adalah kumpulan judul berita *online* berbahasa Indonesia yang menyajikan berbagai peristiwa terkait COVID-19 yang dikenal dengan dataset *Indonesian Corpus for COVID-19 Event Detection* (InaCOVED). Dataset ini diperoleh dari Badan Riset dan Inovasi Nasional yang dapat diakses melalui `https://data.brin.go.id/dataset.xhtml?persistentId=hdl:20.500.12690/RIN/8K14KW`. Proses pengumpulan data dilakukan dengan metode *crawling* yaitu proses pengambilan informasi dari situs web berita *online* yang terdiri dari tujuh portal berita Indonesia seperti “Antara”, “Detik”, “Kompas”, “Merdeka”, “Republika”, “Tempo”, dan “Tirto”. Data dikumpulkan dalam jangka waktu 26 Januari 2020 hingga 24 Mei 2020 dan berjumlah 16.841 data dengan lima variabel yaitu ”_id”, ”portal”, ”published_at_iso” (waktu terbit), ”title” (judul), dan ”event” (jenis peristiwa berita) yang disajikan dalam Tabel 4. sebagai berikut.

Tabel 4. Dataset InaCOVED

_id	portal	published_at_iso	title	event
ObjectId(5ec911d6bda6e09796c064b8)	antara	2018-08-12T13:44:00.000Z	Ronaldo dilaporkan dirawat karena pneumonia	1
ObjectId(5ecd602da19ab66646cd6ca3)	tempo	2018-12-13T07:06:00.000Z	Bayi yang Tak Cukup Mendapat ASI Rentan Pneumonia	0
ObjectId(5eca10a5875c2e623fdfc38e)	kompas	2019-05-07T04:08:00.000Z	[HOAKS] Kurma Timur Tengah Mengandung Virus Corona dari Kelelawar	0
⋮	⋮	⋮	⋮	⋮
ObjectId(5ecdacf980265b0b8b8774ef)	republika	2020-05-26T03:33:00.000Z	Tasikmalaya Dinilai Sudah Lalui Masa Puncak Pandemi Covid-19	1

Penelitian ini berfokus untuk mengklasifikasikan jenis peristiwa COVID-19 dengan menggunakan vaiabel ”title” dan ”event”. Data dikategorikan menjadi 2 kelas yaitu 0 (negatif) yang menunjukkan kelas bukan kejadian *non-event* dan mencakup berita mengenai COVID-19 yang melaporkan peristiwa non-infeksi, seperti tips kesehatan pencegahan COVID-19. Kemudian, kelas 1 (positif) yang menunjukkan kelas kejadian *event* dan mencakup berita laporan peristiwa infeksi COVID-19, seperti kenaikan atau penurunan kasus yang terjadi (Khotimah dkk., 2023). Kelas 0 merupakan kelas mayoritas dengan jumlah data 12.281, sedangkan sisanya merupakan kelas 1 sebanyak 4.540 data yang menjadi kelas minoritas.

3.2.2 Alat Penelitian

3.2.2.1. Spesifikasi Perangkat

Penelitian ini menggunakan laptop merk ASUS dengan model VivoBook ASUS Laptop X409MA. Spesifikasi perangkat tersebut adalah sebagai berikut.

- *Processor* : *Intel(R) Celeron(R)* N4020 CPU @ 1.10GHz
- *Processor speed* : 1101 Mhz
- *Memory* : SSD 256 GB
- RAM : 4 GB
- Sistem Operasi *Windows* 11 (64-bit)

3.2.2.2. *Software* Perangkat

Software Operating System (OS) yang digunakan dalam penelitian ini adalah *Windows 11* dengan bahasa pemrograman *Python* versi 3.11.12. *Google Colab* adalah *platform* yang digunakan untuk menjalankan bahasa pemrograman *Python*. Adapun beberapa *library* yang digunakan dalam bahasa pemrograman *Python* disajikan dalam Tabel 5. sebagai berikut.

Tabel 5. *Library Python*

No	Library	Versi	Tugas
1.	*Googletrans*	4.0.0-rc.1	*Library googletrans* dapat digunakan untuk menerjemahkan teks secara otomatis dari satu bahasa ke bahasa lain menggunakan layanan *Google Translate*.
2.	*Matplotlib*	3.10.0	*Library Matplotlib* dapat membuat beragam grafik untuk visualisasi yang interaktif pada *Python*.
3.	NLTK	3.9.1	*Natural Language Toolkit* (NLTK) merupakan *library* yang dikembangkan khusus untuk tujuan penelitian dan pembelajaran di bidang NLP, pengambilan informasi, *Artificial Intelligence* (AI), dan *Machine Learning* (ML). *Library* ini dapat menguraikan struktur sintaksis dari kalimat tertentu dan melakukan berbagai tugas NLP seperti tokenisasi, *stemming, tagging,* dan tugas lainnya.

No	Library	Versi	Tugas
4.	*NumPy*	2.0.2	*Library Numerical Python (NumPy)* digunakan untuk perhitungan numerik dan *array* secara efisien. *Library* ini memiliki kinerja tinggi untuk komputasi numerik.
5.	*Pandas*	2.2.2	*Library Pandas* digunakan untuk menyiapkan data, memanipulasi data, analisis data, dan menawarkan struktur data seperti *DataFrame*.
6.	*PyTorch*	2.6.0 +cu124	*Library Pytorch* digunakan untuk membangun, melatih, dan menguji model secara efisien, fleksibel, dan mudah digunakan.
7.	*Requests*	2.32.3	*Library requests* digunakan untuk mengirim permintaan HTTP ke *website* atau API dari dalam program *Python*.
8.	*Scikit-learn*	1.6.1	*Library* ini dirancang untuk bekerja sama dengan *NumPy* dan *SciPy*. *Scikit-learn* mencakup berbagai algoritma untuk *unsupervised* and *supervised learning*, serta menyediakan alat untuk evaluasi model dan pra-pemrosesan. Beberapa fitur yang ada dalam *scikit-learn* termasuk pengelompokan, dataset, validasi silang, pemilihan fitur, pengurangan dimensi, penyetelan parameter, dan Ekstraksi fitur.
9.	*Seaborn*	0.13.2	*Library* untuk visualisasi berdasarkan *matplotlib* yang dapat membuat grafik statistik yang menarik, sehingga membuatnya lebih mudah untuk memvisualisasikan hubungan data yang kompleks.
10.	*Stanza*	1.10.1	*Library Stanza* menyediakan berbagai model dan alat untuk melakukan analisis linguistik secara otomatis terhadap teks dalam banyak bahasa. *Library* ini menawarkan *pipeline neural network* yang lengkap, mulai dari tokenisasi, ekspansi multi-kata, *lemmatization, POS tagging*, hingga Named Entity Recognition (NER).
11.	*Transformers*	4.51.3	*Library Transformers* dirancang untuk menangani berbagai tugas NLP secara efektif. Beberapa tugas tersebut diantaranya klasifikasi teks, pemahaman bahasa, terjemahan mesin, ringkasan, aplikasi *multilingual*, dan tugas lainnya.

3.3 Metode Penelitian

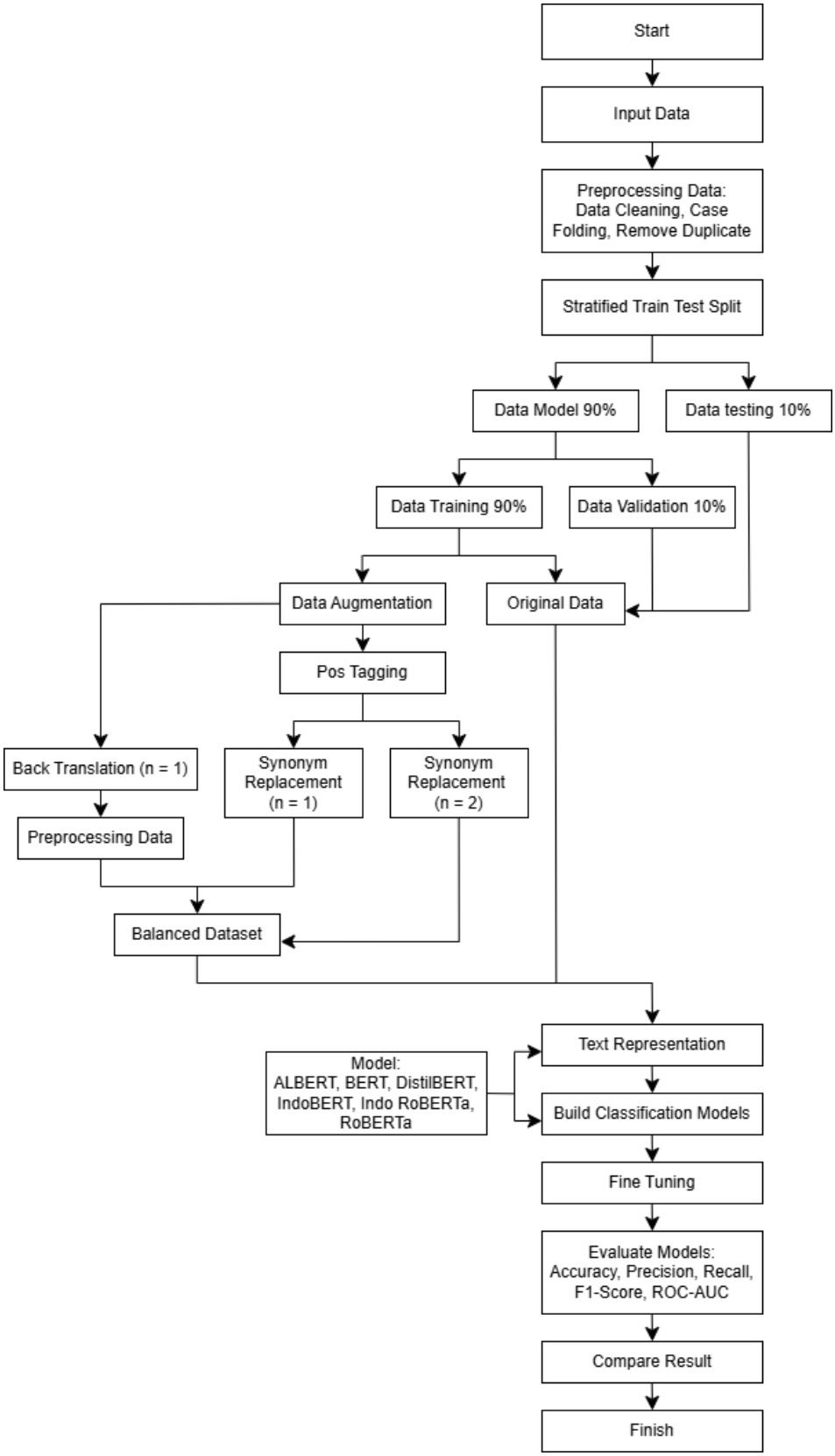
Secara umum, penelitian meliputi beberapa tahap yaitu:

1. Mempersiapkan data InaCOVED yang berisi kumpulan judul berita *online* terkait peristiwa COVID-19.
2. Selanjutnya tahap *text preprocessing* yaitu proses membersihkan data agar terhindar dari hal-hal tidak relevan yang dapat mengganggu proses analisis. Beberapa proses *text preprocessing* yang dilakukan yaitu *data cleaning* berupa penghapusan karakter tab *(remove tab)* dan penghapusan tanda baca atau karakter lain yang tidak digunakan *(remove punctuation)*. *Case folding* dilakukan untuk menyeragamkan format huruf menjadi huruf kecil semua. *Remove duplicate* merupakan proses menghapus data duplikat atau teks yang sama dalam data. Setelah melalui tahap ini data sudah siap untuk diproses ke tahap berikutnya.
3. Membagi data menjadi 2 bagian yaitu data *model* (90%) dan *testing* (10%), kemudian dari 90% data *model* akan dibagi lagi menjadi 2 bagian yaitu *training* (90%) dan *validation* (10%), tahap ini menggunakan metode *stratified train test split* agar pembagian data tetap menjaga proporsi kelas tersebut. Sehingga distribusi akhir pembagian data adalah *training* (81%), *testing* (10%), dan *validation* (9%).
4. Menyimpan hasil *split* data dalam format *pickle*.
5. Menduplikasikan data *training* untuk diproses ke tahap augmentasi.
6. Melakukan augmentasi data dengan teknik *oversampling* yaitu menambah jumlah data berlabel minoritas sehingga jumlahnya sama seperti label mayoritas. Pada penelitian ini, terdapat 2 versi augmentasi yang akan diterapkan yaitu pertama, augmentasi pada data *training* dengan *synonym replacement*. Prosesnya dengan menambah jumlah data training berlabel 1 sebanyak n = 2 menggunakan sinonim teks yang diambil dari kateglo API. Kedua, teknik augmentasi dengan *synonym replacement* dikombinasikan dengan *back translation*. Jumlah n yang digunakan untuk menghasilkan sinonim pada teknik ini adalah n = 1 untuk masing-masing metode, sehingga gabungan dari keduanya akan menghasilkan jumlah data yang cukup untuk menyeimbangkan kelas. Namun, metode *back translation* menghasilkan kalimat dengan tambahan karakter yang kurang relevan, sehingga harus dilakukan *preprocessing* kembali.
7. Menyimpan hasil augmentasi data *training* dengan *synonym replacement* dan hasil augmentasi dengan *synonym replacement + back translation* ke dalam format *pickle*. Sehingga terdapat 3 versi dataset yang akan dianalisis dalam

model yaitu data original, data augmentasi *synonym replacement* (SR), dan data augmentasi *synonym replacement* + *back translation* (SR + BT).

8. Melakukan tokenisasi dengan tipe tokenizer yang disesuaikan dengan model yang akan dibangun. Beberapa model tersebut adalah model ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, dan RoBERTa.
9. Hasil tokenisasi akan dilanjutkan ke proses *embedding* model untuk menghasilkan representasi teks dalam bentuk numerik. Jenis *embedding* akan menyesuaikan dengan model yang digunakan.
10. Membangun model klasifikasi yaitu model *pre-trained* berbasis *transformer* yang telah disempurnakan dan tersedia di *hugging face* serta dapat langsung digunakan seperti ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, dan RoBERTa.
11. Melakukan penyetelan *hyperparameter tuning* untuk mengoptimalkan kinerja model.
12. Mengevaluasi dan membandingkan kinerja masing-masing model.
13. Melakukan *benchmarking* dengan penelitian terdahulu dan membuat keputusan.

Kerangka penelitian untuk klasifikasi judul berita COVID-19 menggunakan *pre-trained transformer models* ditunjukkan pada Gambar 15.



Gambar 15. Kerangka Penelitian

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil dan pembahasan yang telah diuraikan pada Bab IV, diperoleh kesimpulan sebagai berikut:

1. Jumlah *epoch* yang lebih tinggi dapat meningkatkan generalisasi model yang menggunakan data augmentasi. Peningkatan jumlah *epoch* dari 30 ke 50 membawa dampak positif terhadap kinerja setiap model di semua metrik evaluasi dengan peningkatan akurasi sebesar 0,20% - 1,19%, presisi 0,21% - 1,36%, *recall* 0,20% - 1,19%, *F1-score* 0,21% - 1,17%, dan nilai ROC-AUC 0,36% - 2,16%.
2. Pengaplikasian *fine-tuning* pada 6 model *pre-trained transformer* mampu meningkatkan generalisasi model dalam klasifikasi.
3. Model pada data augmentasi berhasil meningkatkan *recall* hingga 14,67%, dan *F1-score* hingga 5,37% pada kelas minoritas. Peningkatan *recall* menunjukkan bahwa model semakin mampu mengenali lebih banyak data positif dari kelas minoritas, yang sebelumnya cenderung terabaikan. Sementara itu, kenaikan *F1-score* menandakan bahwa peningkatan *recall* tidak mengorbankan presisi secara drastis, sehingga keseimbangan antara kedua metrik tetap terjaga.
4. Metode *Synonym Replacement (SR)* berhasil meningkatkan rata-rata nilai akurasi hingga 1,39%, presisi hingga 1,93%, recall hingga 1,39%, F1-score hingga 1,88%, dan ROC-AUC hingga 5,58%. Sedangkan metode *Synonym Replacement + Back Translation (SR + BT)* berhasil meningkatkan rata-rata nilai akurasi hingga 1,45%, presisi hingga 1,96%, recall hingga 1,45%, F1-score hingga 1,94%, dan ROC-AUC hingga 5,54%. Meskipun terdapat beberapa model yang mengalami penurunan pada satu atau lebih metrik evaluasi, secara keseluruhan peningkatan yang diperoleh terutama pada ROC-AUC menunjukkan bahwa augmentasi mampu memperbaiki kemampuan model dalam membedakan antar kelas.

5. Model IndoBERT dengan skenario data augmentasi SR + BT menghasilkan kinerja terbaik dan konsisten secara keseluruhan dengan nilai akurasi 95,25%, presisi 95,24%, recall 95,25%, F1-score 95,25%, dan ROC-AUC 93,90%. Model IndoBERT unggul dibandingkan model lainnya dikarenakan model IndoBERT telah dilatih khusus pada korpus besar berbahasa Indonesia. Selain itu, dataset yang digunakan dalam penelitian ini adalah data judul berita *online* berbahasa Indonesia sehingga model IndoBERT sangat sesuai untuk diaplikasikan dalam tugas klasifikasi yang berbahasa Indonesia.
6. Augmentasi data berhasil meningkatkan kinerja model ALBERT, BERT, DistilBERT, IndoBERT, Indo RoBERTa, dan RoBERTa. Model berhasil mendapat manfaat dari data latih yang telah diseimbangkan. Model dengan distribusi data latih seimbang mampu mengidentifikasi kelas minoritas pada data InaCOVED dengan lebih baik dibandingkan model yang menggunakan data tidak seimbang.
7. Temuan ini dapat digunakan sebagai sistem klasifikasi otomatis untuk data baru yang belum pernah dilihat.

5.2 Saran

Pada penelitian ini, penulis berfokus membahas pengaruh augmentasi data dalam dataset yang tidak seimbang dan juga mengembangkan berbagai model *pre-trained transformer* untuk tugas klasifikasi. Oleh sebab itu, adapun saran yang diberikan untuk penelitian selanjutnya yaitu sebagai berikut:

1. Menambahkan proses validasi kualitas augmentasi, seperti penggunaan *cosine similarity* serta penyaringan berbasis *threshold* atau anotator, karena hasil augmentasi baik dari *synonym replacement* maupun *back translation* masih berpotensi mengandung *noise* yang dapat menurunkan kualitas pelatihan model. Oleh karena itu, evaluasi ulang terhadap kalimat hasil augmentasi perlu dilakukan sebelum digunakan dalam pelatihan model.
2. Mengembangkan teknik augmentasi lain seperti *generate text* yang dapat menghasilkan kalimat baru yang lebih bervariasi.
3. Menggunakan perangkat dengan spesifikasi yang cukup tinggi seperti *processor* Intel@ Core i5 atau Intel@ Core i7 dan RAM 8 GB, sistem operasi minimum 64 *bit* karena proses *running* dengan model *pre-trained transformer* membutuh waktu komputasi yang cukup lama.

DAFTAR PUSTAKA

Agarap, A. F. 2018. Deep Learning using Rectified Linear Units (ReLU). *arXiv preprint arXiv:1803.08375*

Babaali, K. O., Zigh, E., Djebbouri, M., Chergui, O. 2022. A new approach for road extraction using data augmentation and semantic segmentation. *Indonesian Journal of Electrical Engineering and Computer Science*. **28**(3): 1493-1501.

Bekkar, M., Djemaa, H. K., Alitouche, T. A. 2013. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*. **3**(10).

Benselin, J.C. dan Ragsdell, G. 2016. Information overload : the differences that age makes. *Journal of Librarianship and Information Science*. **48**(3): 284-297.

Cochran, K., Cohn, C., Hutchins, N., Biswas, G., Hastings, P. 2022. Improving automated evaluation of formative assessments with text data augmentation. *International Conference on Artificial Intelligence in Education*. hlm, 390-401.

Dai, L., Zhou, M., Liu, H. 2024. Recent Applications of Convolutional Neural Networks in Medical Data Analysis. *IGI Global Scientific Publishing*. hlm, 119-131.

Devi, D., Biswas, S. K., Purkayastha, B. 2020. A review on solution to class imbalance problem: Undersampling approaches. *2020 international conference on computational performance evaluation*. hlm, 626-631.

Devlin, J., Chang, M. W., Lee, K., Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*. hlm, 4171-4186.

Elbasani, E., Njimbouom, S. N., Oh, T. J., Kim, E. H., Lee, H., Kim, J. D. 2021. GCRNN: graph convolutional recurrent neural network for compound–protein interaction prediction. *BMC bioinformatics*. **22**(5): 616-628.

Fausett, L. 1994. *Fundamentals of Neural Networks : Architectures, Algorithms, and Applications*. Prentice Hall, New Jersey.

Feldman, R., Sanger, J. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., Japkowicz, N. 2024. The class imbalance problem in deep learning. *Machine Learning*. **113**(7): 4845-4901.

Gonzalez-Carvajal, S., Garrido-Merchan, E. C. 2020. Comparing BERT against traditional machine learning text classification. *Journal of Computational and Cognitive Engineering*. hlm, 1-10.

Gupta, P., Gandhi, S., Chakravarthi, B. R. 2021. Leveraging transfer learning techniques-bert, roberta, albert and distilbert for fake review detection. *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*. hlm, 75-82.

Han, J., Kamber, M. 2006. *Data Mining: Concepts and Techniques, Second Edition*. Morgan kaufmann Publishers.

Hancock, J. T., Wang, H., Khoshgoftaar, T. M., Liang, Q. 2024. Data reduction techniques for highly imbalanced medicare Big Data. *Journal of Big Data*. **11(1)**: 8.

Herrera, F., Charte, F., Rivera, A. J., Del Jesus, M. J., Herrera, F., Charte, F., ... del Jesus, M. J. 2016. *Multilabel classification*. Springer International Publishing. hlm, 17-31.

Jungiewicz, M., Smywinski-Pohl, A. 2019. Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science*. **20**(1): 57-83.

Khan, N., Siddiqui, B. N., Khan, N., Ahmad, Z., Ismail, S., Javed, H. H., ... Kasi, A. K. 2020. Mass media role in agricultural and rural development. *International Journal of Advanced Research in Biological Sciences*. **7**(4): 199-209.

Khotimah, P. H., Arisal, A., Rozie, A. F., Nugraheni, E., Riswantini, D., Suwarningsih, W., ... Purwarianti, A. 2023. Monitoring Indonesian online news for COVID-19 event detection using deep learning. *International Journal of Electrical Computer Engineering*. **13**(1): 2088-8708.

Kingma, D. P., Ba, J. L. 2015. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations San Diego*. hlm, 1–15.

Krstajic, M., Najm-Araghi, M., Mansmann, F., Keim, D. A. 2013. Story tracker: Incremental visual text analytics of news story development. *Information Visualization*. **12**(3-4): 308-323.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Madabushi, H. T., Kochkina, E., Castelle, M. 2020. Cost-sensitive BERT for generalisable sentence classification with imbalanced data. *arXiv preprint arXiv:2003.11563*.

Nabiilah, G. Z., Prasetyo, S. Y., Izdihar, Z. N., Girsang, A. S. 2023. BERT base model for toxic comment analysis on Indonesian social media. *Procedia Computer Science*. **216**: 714-721.

Nabiilah, G. Z., Suhartono, D. 2023. Personality Classification Based on Textual Data using Indonesian Pre-Trained Language Model and Ensemble Majority Voting. *Revue d'Intelligence Artificielle*. **37**(1): 73-81.

Ningsih, F. S. S., Khotimah, P. H., Arisal, A., Rozie, A. F., Munandar, D., Riswantini, D., ... Kurniasari, D. 2022. Synonym-based Text Generation in Restructuring Imbalanced Dataset for Deep Learning Models. *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS)*. hlm, 1-6.

Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S. 2018. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C. D. 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Rahma, I. A., Suadaa, L. H. 2023. Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer*. **10**(6): 1329-1340.

Rasamoelina, A. D., Adjailia, F., Sinčák, P. 2020. A review of activation function for artificial neural network. *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. hlm, 281-286.

Rohman, AN, Utami, E., Raharjo, S. 2019. Deteksi Kondisi Emosional di Media Sosial Menggunakan Pendekatan Lexicon dan Natural Language Processing. *Jurnal Eksplorasi Informatika*. **9**(1): 70-76.

Romli, A. S. M. 2018. *Jurnalistik online: Panduan mengelola media online*. Nuansa Cendekia.

Sanh, V., Debut, L., Chaumond, J., Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shahrzadi, L., Mansouri, A., Alavi, M., Shabani, A. 2024. Causes, consequences, and strategies to deal with information overload: A scoping review. *International Journal of Information Management Data Insights*. **4**(2): 100261.

Sharifirad, S., Jafarpour, B., Matwin, S. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. *Proceedings of the 2nd workshop on abusive language online (ALW2)*. hlm, 107-114.

Sirusstara, J., Alexander, N., Alfarisy, A., Achmad, S., Sutoyo, R. 2022. Clickbait headline detection in indonesian news sites using robustly optimized bert pre-training approach (roberta). *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. hlm, 1-6.

Vairetti, C., Assadi, J. L., Maldonado, S. 2024. Efficient hybrid oversampling and intelligent undersampling for imbalanced big data classification. *Expert Systems with Applications*. **246**: 123149.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*. **30**: 5998-6008.

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., Lestantyo, P. 2019. Cross-validation metrics for evaluating classification performance on imbalanced data. *2019 international conference on computer, control, informatics and its applications (IC3INA)*. hlm, 14-18.

Widarmanti, T., Widodo, M. P., Ramadhani, D. P., Danlami, M. 2022. Text Emotion Detection: Discover the Meaning Behind YouTube Comments Using Indo RoBERTa. *2022 International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS)*. hlm, 1-6.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., ... Purwarianti, A. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *arXiv preprint arXiv:2009.05387*.

Wirawan, A., Cahyono, H. D. 2023. Easy Data Augmentation in Sentiment Analysis of Cyberbullying. *2023 6th International Conference on Information and Communications Technology (ICOIACT)*. hlm, 443-447.

Xu, C., Coen-Pirani, P., Jiang, X. 2023. Empirical study of overfitting in deep learning for predicting breast Cancer metastasis. *Cancers*. **15**(7) : 1969.

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., Langlotz, C. P. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*. **28**(9) : 1892-1899.

Zheng, H., Wang, X., Huang, Y. H. 2023. Fake news in a time of plague: Exploring individuals' online information management in the COVID-19 era. *Computers in Human Behavior*. **146**: 107790.