BAB 3
METODOLOGI PENELITIAN

3.1 Pengumpulan Data

Dataset dalam penelitian ini dikembangkan menggunakan varian model bahasa besar GPT-4, yakni GPT-4o dan GPT-4o mini, dengan memanfaatkan prompt yang dirancang secara spesifik dan terperinci. Prompt tersebut memberikan instruksi kepada model untuk: *"Pretend you are researching fine-tuning LLM for mixed-language hate speech detection. Generate 50 UNIQUE mixed-language data about hate speech. Ensure the content relates to [Object]. For hate speech labeled data, generate distinct hate speech and differentiate it from negative sentiment. Hate speech should belittle, discriminate against, or incite violence based on protected characteristics."* Instruksi ini memastikan bahwa data yang dihasilkan relevan, jelas terlabel, dan mampu membedakan secara halus antara ujaran kebencian dengan sentimen negatif.

Model menjalankan prompt tersebut dalam 600 sesi, menghasilkan data dalam tiga kategori bahasa: bahasa Indonesia, bahasa Inggris, dan campuran (mixed-language). Setiap kategori dirancang dengan proporsi yang seimbang, yaitu 50,2% untuk ujaran kebencian dalam bahasa Indonesia, 49,9% dalam bahasa Inggris, dan 50,4% dalam bahasa campuran. Secara keseluruhan, dataset terdiri dari 9.783 entri dalam bahasa Indonesia, 9.968 entri dalam bahasa Inggris, dan 7.835 entri dalam bahasa campuran, dengan rasio 50:50 antara data ujaran kebencian dan bukan kebencian.

Pertimbangan etika menjadi prioritas utama mengingat sifat konten yang sensitif. Seluruh data berasal dari sumber yang tersedia secara publik, dan informasi yang dapat mengidentifikasi individu telah dihapus untuk menjaga anonimitas serta mematuhi standar etika penelitian. Selain itu, guna memperkaya variasi linguistik tanpa mengubah makna inti teks, dilakukan proses augmentasi data menggunakan Natural Language Toolkit (NLTK). Proses ini mencakup teknik penggantian sinonim, di mana kata benda, kata sifat, dan kata kerja penting dalam kalimat diidentifikasi lalu digantikan dengan sinonim yang sesuai.

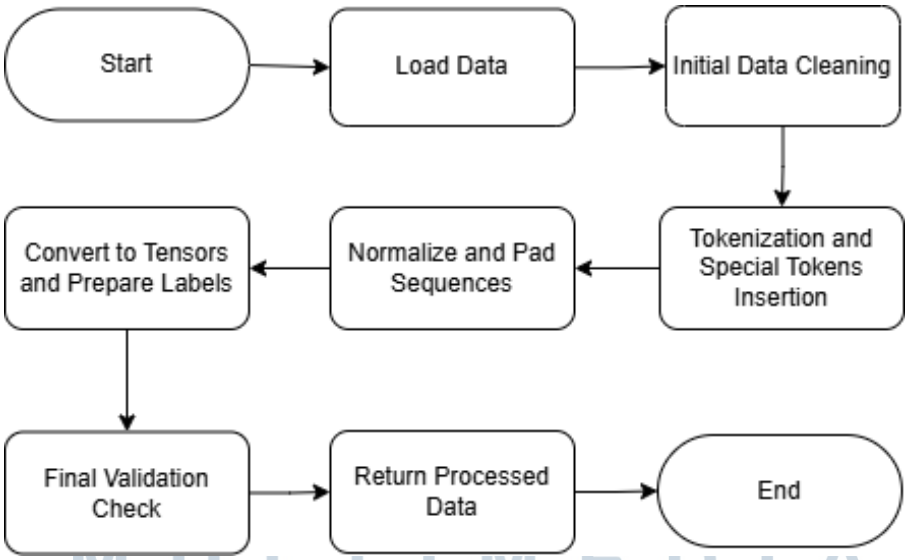
Pemilihan sinonim dilakukan secara cermat agar tetap mempertahankan konteks dan sentimen asli dalam teks. Untuk menjaga integritas data, dalam setiap kalimat biasanya hanya satu kata utama yang digantikan. Pendekatan sistematis

ini memungkinkan ekspansi dataset dengan cakupan ekspresi linguistik yang lebih luas, yang penting untuk pelatihan model pembelajaran mesin yang lebih tangguh dan sensitif terhadap konteks.

Dataset akhir terdiri dari 30.000 entri dan telah didokumentasikan secara menyeluruh. Seluruh dataset tersedia secara publik melalui repositori GitHub. Ketersediaan ini mendukung prinsip transparansi dan reproduktabilitas, serta memberikan wawasan lebih lanjut tentang penggunaan NLTK dalam pengayaan dataset, sehingga dapat dimanfaatkan dalam penelitian lanjutan di bidang ini.

3.2 Pra-pemrosesan Data

Pra-pemrosesan data merupakan tahapan awal yang esensial dalam proses pembangunan sistem klasifikasi berbasis pembelajaran mesin. Tahapan ini bertujuan untuk memastikan bahwa data mentah yang diperoleh dari sumber eksternal telah berada dalam kondisi yang layak dan sesuai untuk digunakan oleh model. Kualitas hasil akhir dari suatu model sangat bergantung pada kualitas data yang digunakan dalam pelatihannya, sehingga proses pra-pemrosesan harus dilakukan secara sistematis dan menyeluruh.



Gambar 3.1. Pra-pemrosesan data

Seperti yang ditunjukkan pada Gambar 3.1, proses pra-pemrosesan data dimulai dengan memuat dataset dasar dari penyimpanan, memastikan format data telah sesuai dengan penamaan kolom yang diperlukan, yaitu `sentence` untuk teks dan `label` untuk target klasifikasinya. Proses awal ini berfokus pada *pembersihan*

data (data cleaning) untuk meningkatkan integritas data, seperti menangani nilai yang hilang, spasi berlebih, kesalahan encoding, dan nilai pencilan (outliers). Langkah ini penting untuk memastikan landasan data yang bersih dan seragam sebelum proses pelatihan model dilakukan.

Pada tahap *tokenisasi dan penyisipan token khusus*, setiap entri teks diubah menjadi token-token diskrit menggunakan tokenizer yang kompatibel dengan arsitektur model BLOOM. Token khusus `[CLS]` disisipkan di awal setiap entri sebagai penanda awal yang membantu model dalam mengagregasi informasi sekuensial untuk keperluan klasifikasi. Sementara itu, token `[SEP]` digunakan untuk membatasi bagian teks atau menunjukkan akhir entri, sehingga model dapat lebih mudah membedakan segmen-segmen teks secara efektif.

Langkah berikutnya adalah *normalisasi dan padding sekuens*, di mana seluruh entri diseragamkan agar memiliki panjang sekuens yang konsisten. Ini dilakukan dengan memotong teks yang terlalu panjang dan menambahkan padding pada teks yang terlalu pendek agar sesuai dengan panjang maksimum tertentu. Panjang ini ditentukan berdasarkan distribusi panjang teks pada dataset atau mengikuti panjang maksimum standar yang umum digunakan untuk model BLOOM. Standarisasi ini sangat penting agar jaringan saraf menerima input dengan format yang konsisten, sehingga model dapat mempelajari data secara efisien.

Setelah normalisasi dan padding selesai, data diubah menjadi format tensor, yang merupakan format yang dibutuhkan untuk pelatihan jaringan saraf. Proses ini juga mencakup konversi label menjadi tensor agar sejajar dengan input model. Sebelum digunakan pada proses pelatihan, validasi akhir dilakukan untuk memastikan tidak ada lagi masalah integritas data yang tersisa. Pra-pemrosesan ini, seperti ditunjukkan pada Gambar 3.1, dirancang secara sistematis untuk memastikan bahwa dataset telah dikonfigurasi secara optimal, sehingga mampu mendukung proses pelatihan dan pengujian model BLOOM secara efisien dan efektif.

3.3 Fine-Tuning

Model BLOOM-560m yang telah dilatih sebelumnya, awalnya dikembangkan untuk memahami dan memproses berbagai informasi linguistik, disesuaikan kembali (fine-tuned) agar kemampuannya dapat difokuskan pada dataset khusus yang terdiri dari 30.000 entri. Proses fine-tuning ini sangat penting untuk meningkatkan performa model dalam tugas-tugas klasifikasi ujaran

kebencian berdasarkan data tersebut.

3.3.1 Pembagian Data

Dataset dibagi secara metodis ke dalam tiga bagian: pelatihan, validasi, dan pengujian, dengan rasio masing-masing 70:20:10. Ini menghasilkan 21.000 entri untuk pelatihan, 6.000 untuk validasi, dan 3.000 untuk pengujian. Pembagian ini mengacu pada praktik terbaik dalam pembelajaran mesin untuk mengoptimalkan proses pembelajaran model dan efektivitas validasinya.

Data Pelatihan (70% – 21.000 entri) digunakan untuk menyesuaikan dan menyempurnakan bobot parameter dalam model. Volume data yang besar ini memungkinkan model mempelajari berbagai pola dan nuansa linguistik, sehingga membentuk kemampuan prediktif yang lebih kuat dan komprehensif.

Data Validasi (20% – 6.000 entri) berfungsi untuk memantau performa model selama proses fine-tuning, memastikan bahwa penyesuaian model memberikan peningkatan yang nyata terhadap data yang menyerupai kondisi nyata. Selain itu, dataset ini digunakan untuk penyetelan hiperparameter, seperti learning rate, batch size, dan arsitektur model, tanpa risiko overfitting karena data ini tidak digunakan dalam pelatihan langsung.

Data Pengujian (10% – 3.000 entri) disediakan untuk evaluasi akhir guna menilai kemampuan generalisasi model terhadap data yang benar-benar baru. Proses ini memberikan tolok ukur yang objektif untuk menilai efektivitas dan kesiapan model dalam menghadapi kondisi dunia nyata.

Pembagian yang sistematis ini memastikan bahwa proses pelatihan, penyetelan, dan evaluasi berjalan optimal dan tidak saling memengaruhi. Ini penting untuk menghasilkan model yang seimbang antara performa dan generalisasi.

3.3.2 Pelatihan dan Validasi

Model BLOOM-560m disesuaikan menggunakan algoritma optimisasi Adam, yang terbukti efektif dalam menangani dataset besar dan cocok untuk skenario fine-tuning. Optimizer ini memiliki keunggulan dalam menangani gradien yang jarang muncul serta kemampuan adaptasi terhadap learning rate.

Tabel 3.1. Hasil metrik Fine-Tuning selama 11 epoch

Epoch	Train Loss	Validation Loss	Train Accuracy	Validation Accuracy	Train Precision	Validation Precision	Training Recall	Validation Recall	Training F1	Validation F1
1	0.1361	0.0663	0.9712	0.9861	0.9708	0.9760	0.9715	0.9967	0.9712	0.9862
2	0.0530	0.0914	0.9896	0.9854	0.9896	0.9933	0.9895	0.9775	0.9896	0.9853
3	0.0090	0.0460	0.9978	0.9891	0.9977	0.9894	0.9979	0.9887	0.9978	0.9891
4	0.0018	0.0455	0.9998	0.9905	0.9997	0.9930	0.9998	0.9881	0.9998	0.9905
5	0.0007	0.0456	0.9999	0.9899	0.9998	0.9875	1.0000	0.9924	0.9999	0.9899
6	**0.0005**	**0.0443**	**0.9999**	**0.9904**	**0.9998**	**0.9888**	**1.0000**	**0.9911**	**0.9999**	**0.9904**
7	0.0004	0.0446	1.0000	0.9904	0.9999	0.9888	1.0000	0.9921	1.0000	0.9904
8	0.0004	0.0448	1.0000	0.9905	0.9999	0.9888	1.0000	0.9924	1.0000	0.9906
9	0.0004	0.0448	1.0000	0.9905	0.9999	0.9888	1.0000	0.9924	1.0000	0.9906
10	0.0004	0.0448	1.0000	0.9905	0.9999	0.9888	1.0000	0.9924	1.0000	0.9906
11	0.0004	0.0448	1.0000	0.9905	0.9999	0.9888	1.0000	0.9924	1.0000	0.9906

Pada tahap ini, digunakan learning rate sebesar `5e-5` dan ukuran batch sebanyak 16. Nilai-nilai ini umum digunakan dalam praktik fine-tuning karena memberikan keseimbangan antara kecepatan konvergensi dan pencegahan overfitting. Learning rate yang moderat memungkinkan pembaruan parameter yang stabil, sementara ukuran batch kecil memberikan pembaruan bobot yang lebih sering dan presisi.

Selama proses pelatihan, performa model dimonitor secara berkala menggunakan data validasi. Mekanisme *early stopping* diterapkan untuk menghentikan pelatihan jika tidak terdapat peningkatan nilai loss validasi selama lima epoch berturut-turut. Strategi ini bertujuan untuk mencegah pelatihan berlebihan serta menghemat waktu dan sumber daya.

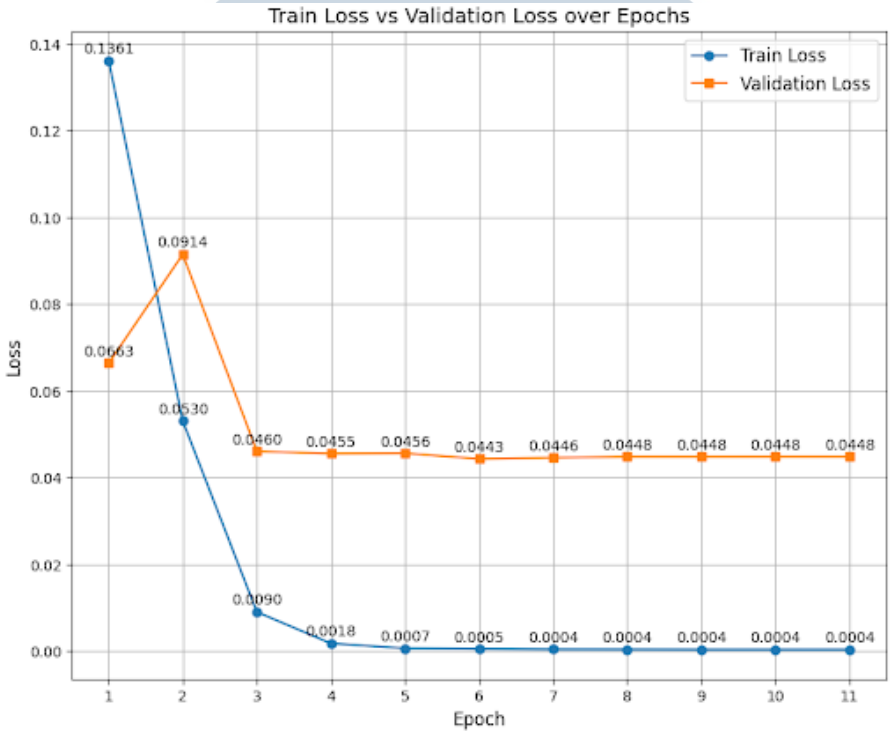
Tujuan utama dari fine-tuning ini adalah untuk meminimalkan loss validasi, yang menjadi indikator utama keseimbangan antara akurasi dan kemampuan generalisasi model. Hasil pelatihan dan validasi didokumentasikan secara menyeluruh menggunakan metrik seperti akurasi, presisi, recall, dan F1-score, sebagaimana ditampilkan dalam Tabel 3.1.

Pada epoch ke-6, tercapai nilai loss validasi terendah yaitu 0,0443, bersamaan dengan peningkatan signifikan pada metrik performa lainnya: akurasi (99,04%), presisi (99,98%), recall (99,11%), dan F1-score (99,04%). Setelah titik ini, tidak terdapat peningkatan lebih lanjut, sehingga pelatihan dihentikan secara otomatis. Keputusan ini menunjukkan bahwa model telah mencapai performa optimal dalam belajar dan generalisasi.

Fokus pada nilai loss validasi sebagai tolok ukur utama dalam strategi early stopping menjamin bahwa peningkatan performa yang dicapai benar-benar mencerminkan kemampuan model untuk menangani data baru, bukan sekadar penyesuaian terhadap data pelatihan.

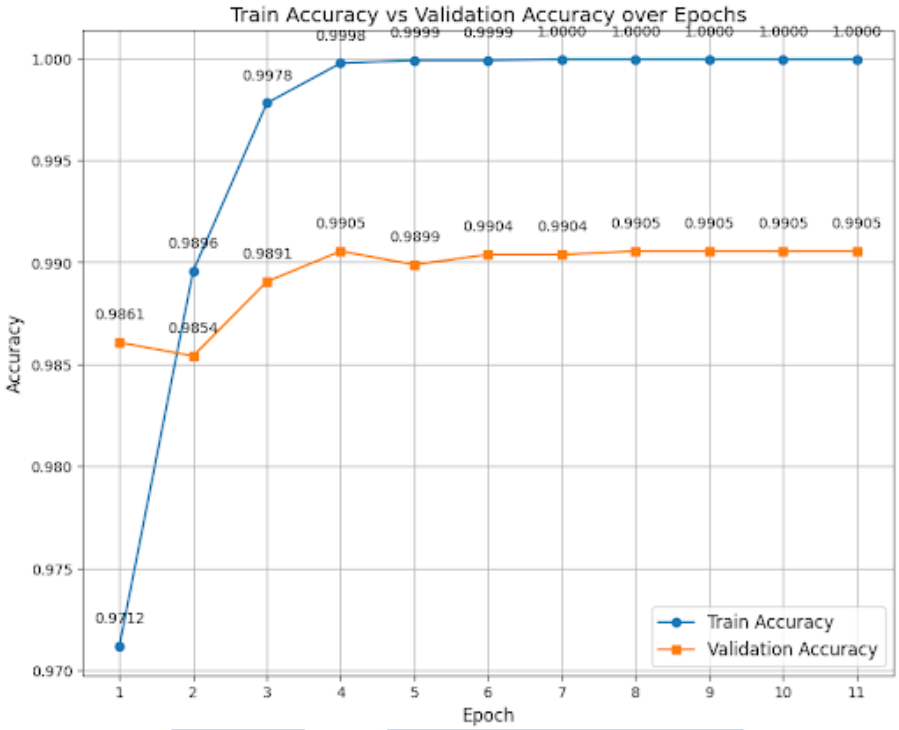
Visualisasi tren metrik pelatihan dan validasi disajikan dalam beberapa grafik berikut, yang menunjukkan perkembangan model selama proses fine-tuning. Grafik-grafik ini tidak hanya menggambarkan arah perubahan dari waktu ke waktu, tetapi juga memberikan pandangan menyeluruh terhadap dinamika pembelajaran model, termasuk fase awal eksplorasi, titik konvergensi, hingga potensi stabilisasi atau overfitting. Setiap metrik—mulai dari loss hingga F1-score—diilustrasikan secara terpisah untuk memberikan penekanan pada aspek performa yang berbeda. Pola konvergensi yang serupa antara metrik pelatihan dan validasi menjadi indikator kuat bahwa model tidak hanya belajar dengan baik dari data yang tersedia, tetapi juga mempertahankan kemampuannya dalam mengenali pola pada data

baru. Dengan demikian, grafik-grafik ini berfungsi sebagai alat evaluasi visual yang penting untuk mendukung keputusan dalam proses tuning, khususnya dalam penerapan strategi seperti early stopping.



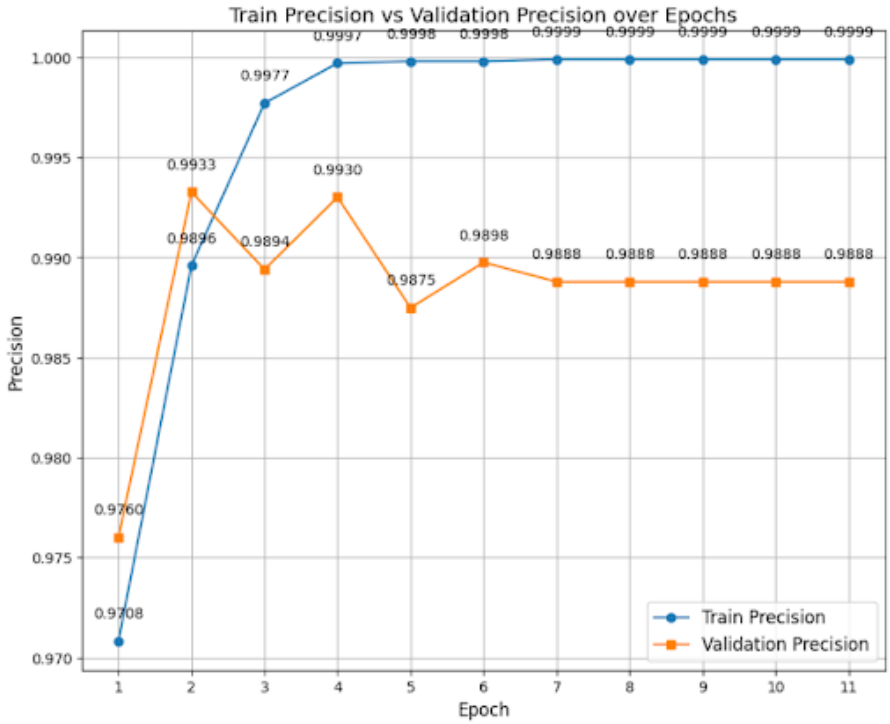
Gambar 3.2. Loss pelatihan dan validasi selama 11 epoch

Gambar 3.2 memperlihatkan tren nilai loss pada data pelatihan dan validasi selama 11 epoch untuk model BLOOM-560m. Loss pelatihan dimulai cukup tinggi pada 0,1361 di epoch ke-1, menurun tajam menjadi 0,0530 pada epoch ke-2, dan terus menurun stabil hingga mencapai 0,0005 pada epoch ke-6, lalu stabil di 0,0004 untuk epoch selanjutnya. Sebaliknya, loss validasi diawali pada 0,0663, naik ke puncaknya di 0,0914 pada epoch ke-2, dan kemudian menurun secara konsisten hingga mencapai titik terendah di 0,0443 pada epoch ke-6 yang diidentifikasi sebagai epoch dengan performa terbaik. Setelah epoch ke-6, nilai loss validasi cenderung stabil di sekitar 0,0448, menunjukkan bahwa tidak ada peningkatan signifikan dalam kemampuan generalisasi model.



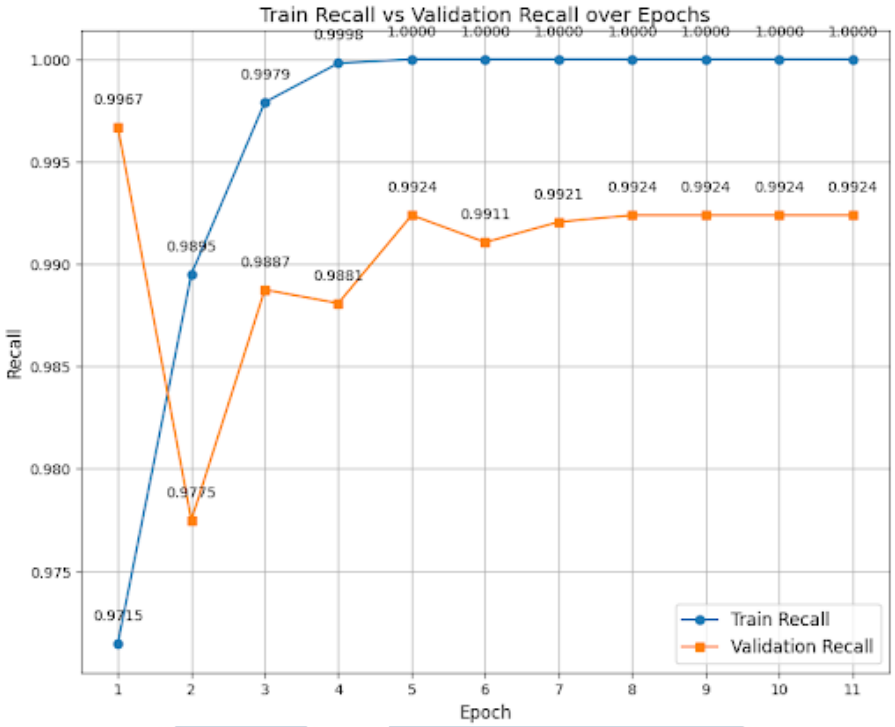
Gambar 3.3. Akurasi pelatihan dan validasi selama 11 epoch

Gambar 3.3 menampilkan tren akurasi pelatihan dan validasi. Akurasi pelatihan dimulai dari 97,12% di epoch pertama, meningkat tajam menjadi 98,96% di epoch ke-2, dan terus meningkat hingga mencapai 99,99% pada epoch ke-6, serta stabil di angka 100% mulai epoch ke-7. Akurasi validasi dimulai pada 98,61% di epoch pertama, sedikit menurun menjadi 98,54% di epoch ke-2, kemudian meningkat secara konsisten dan mencapai puncak di 99,05% pada epoch ke-6. Tren ini menunjukkan konvergensi antara performa model pada data pelatihan dan validasi, dengan epoch ke-6 sebagai titik keseimbangan terbaik.



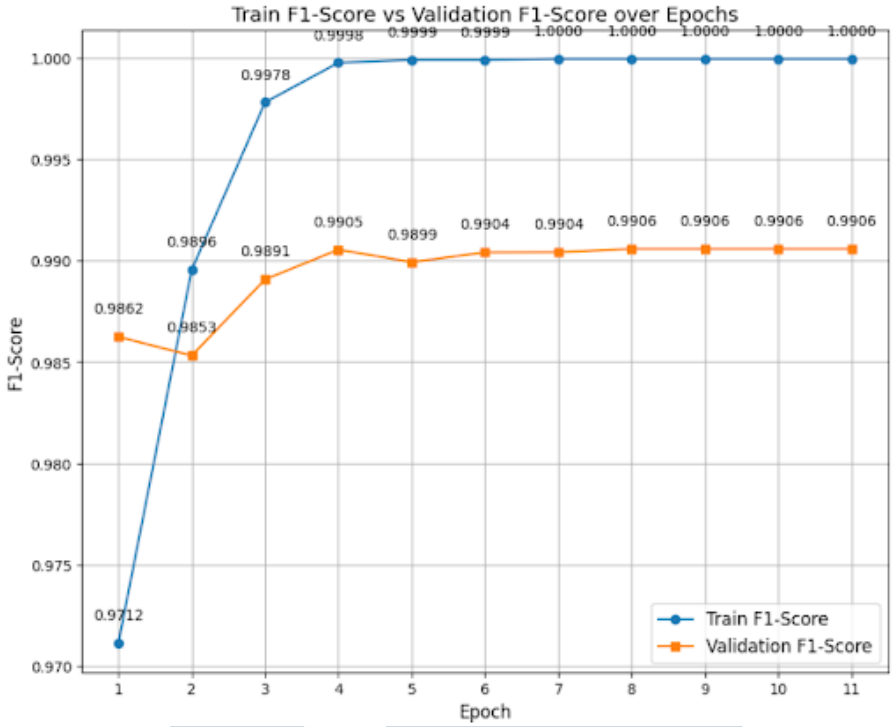
Gambar 3.4. Presisi pelatihan dan validasi selama 11 epoch

Gambar 3.4 menunjukkan perkembangan presisi pada data pelatihan dan validasi. Presisi pelatihan dimulai pada 97,08% di epoch pertama, meningkat cepat menjadi 99,77% di epoch kedua, dan mencapai 99,99% di epoch ke-6, lalu stabil. Presisi validasi dimulai pada 97,60% dan meningkat menjadi 99,33% pada epoch kedua, dengan puncak 99,30% di epoch ke-4. Setelah itu, terjadi sedikit fluktuasi, turun ke 98,75% di epoch ke-5, dan stabil di 98,88% mulai epoch ke-6. Hal ini menunjukkan bahwa presisi model terhadap data baru tetap stabil dan konsisten setelah fase awal pelatihan.



Gambar 3.5. Recall pelatihan dan validasi selama 11 epoch

Gambar 3.5 menggambarkan tren recall. Recall pelatihan dimulai dari 97,15% di epoch pertama, meningkat drastis menjadi 98,95% di epoch kedua, 99,79% di epoch ketiga, dan mencapai 100% di epoch kelima, lalu stabil. Recall validasi dimulai sangat tinggi di 99,67% pada epoch pertama, menurun ke 97,75% di epoch kedua, lalu meningkat dan mencapai puncak di 99,24% pada epoch kelima dan stabil setelahnya. Ini menunjukkan performa model yang konsisten dalam mengenali data positif secara akurat setelah awal yang fluktuatif.



Gambar 3.6. F1-Score pelatihan dan validasi selama 11 epoch

Gambar 3.6 menunjukkan perkembangan nilai F1-Score. F1-Score pelatihan dimulai pada 97,12% di epoch pertama, meningkat ke 98,96% di epoch kedua, 99,78% di epoch ketiga, dan mencapai 100% mulai epoch ke-7. F1-Score validasi diawali pada 98,62%, sedikit menurun ke 98,53% di epoch kedua, lalu meningkat dan mencapai titik tertinggi di 99,06% pada epoch kedelapan, kemudian stabil. Hal ini mengindikasikan stabilitas performa model dalam menjaga keseimbangan antara presisi dan recall.

3.4 Penyetelan Hiperparameter

Pada tahap penyetelan hiperparameter model BLOOM-560m, digunakan pendekatan *grid search* untuk menyempurnakan performa model secara sistematis melalui eksplorasi berbagai kombinasi ukuran batch dan nilai *learning rate*. Ukuran batch yang diuji adalah 8 dan 16, sementara nilai *learning rate* yang digunakan berkisar antara `1e-5` hingga `5e-5`. Pemilihan rentang ini mengacu pada praktik umum dalam proyek pembelajaran mesin berskala besar, di mana kombinasi tersebut terbukti memberikan keseimbangan antara efisiensi komputasi dan akurasi model.

Tabel 3.2. Hasil Penyesuaian Hyperparameter Selama Beberapa Epoch untuk Berbagai Ukuran Batch dan Learning Rate

LR	Batch Size	Best Epoch	Train Loss	Validation Loss	Train Accuracy	Validation Accuracy	Train Precision	Validation Precision	Training Recall	Validation Recall	Training F1	Validation F1
5e-5	8	17	**0.0112**	**0.0007**	0.9988	**0.9997**	0.9988	**0.9997**	0.9988	**0.9997**	0.9988	**0.9997**
4e-5	8	19	0.0063	0.0008	0.9988	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3e-5	8	6	0.1260	0.0151	0.9707	0.9944	0.9707	0.9944	0.9707	0.9944	0.9707	0.9944
2e-5	8	9	0.0266	0.0056	0.9925	0.9980	0.9926	0.9980	0.9925	0.9980	0.9925	0.9980
1e-5	8	2	2.5364	0.2770	0.8634	0.9413	0.8634	0.9458	0.8633	0.9414	0.8634	0.9412
5e-5	16	9	0.0092	0.0033	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992	0.9992
4e-5	16	11	0.01207	0.0076	0.9985	0.9985	0.9985	0.9985	0.9985	0.9985	0.9985	0.9985
3e-5	16	7	0.0672	0.0121	0.9816	0.9983	0.9816	0.9983	0.9816	0.9983	0.9816	0.9983
2e-5	16	16	0.0100	0.0047	0.9969	0.9983	0.9969	0.9984	0.9969	0.9983	0.9969	0.9983
1e-5	16	7	0.0425	0.0170	0.9862	0.9957	0.9862	0.9957	0.9862	0.9957	0.9862	0.9957

Mekanisme *early stopping* tetap digunakan pada fase ini untuk mencegah overfitting, dengan menghentikan pelatihan apabila tidak ada peningkatan loss validasi setelah lima epoch berturut-turut. Mekanisme ini menjadi penting untuk semua kombinasi batch dan learning rate karena memastikan proses pelatihan tetap optimal tanpa membuang sumber daya secara berlebihan.

Setiap konfigurasi diperbolehkan untuk dilatih hingga maksimal 20 epoch. Batasan ini ditetapkan untuk menjaga efisiensi pelatihan dengan tetap memungkinkan kedalaman pembelajaran yang memadai. Dengan membatasi jumlah epoch, risiko pembelajaran berlebihan atau peningkatan performa yang tidak signifikan dapat diminimalkan, terutama ketika berhadapan dengan dataset besar dan arsitektur model yang kompleks.

Jika terjadi peningkatan pada loss validasi selama pelatihan, bobot model pada epoch tersebut disimpan sebagai *checkpoint* terbaik. Dengan cara ini, konfigurasi akhir model akan merujuk pada kondisi terbaik yang tercapai selama penyetelan. Hal ini menjamin model yang dihasilkan tidak hanya akurat, tetapi juga efisien dan mampu digunakan secara praktis dalam skenario dunia nyata.

Dengan memanfaatkan model BLOOM-560m hasil fine-tuning sebagai titik awal, eksplorasi variasi hiperparameter dilakukan untuk lebih mengoptimalkan model terhadap tugas klasifikasinya. Hasil dari proses ini memberikan wawasan penting mengenai pengaruh ukuran batch dan learning rate terhadap performa model, dan menjadi dasar dalam memilih kombinasi konfigurasi terbaik yang mampu memberikan keseimbangan antara akurasi, efisiensi, dan stabilitas.

Berdasarkan hasil pada Tabel 3.2, konfigurasi dengan learning rate `5e-5`, batch size 8, dan epoch terbaik di epoch ke-17 memberikan performa tertinggi. Konfigurasi ini menghasilkan loss validasi sebesar 0,0007 dan loss pelatihan 0,0112. Selain itu, model mencapai metrik validasi yang sangat tinggi: akurasi, presisi, recall, dan F1-score masing-masing sebesar 99,97%.

Beberapa konfigurasi lain menunjukkan performa yang kompetitif, namun tidak melebihi performa optimal konfigurasi terbaik. Learning rate yang lebih tinggi seperti `4e-5` cenderung menghasilkan loss validasi yang sedikit lebih besar, sedangkan batch size yang lebih besar seperti 16 memerlukan lebih sedikit epoch untuk mencapai performa terbaik. Temuan ini menegaskan pentingnya penyetelan kombinasi learning rate dan batch size untuk mendapatkan keseimbangan terbaik antara efisiensi pelatihan dan kemampuan generalisasi model.

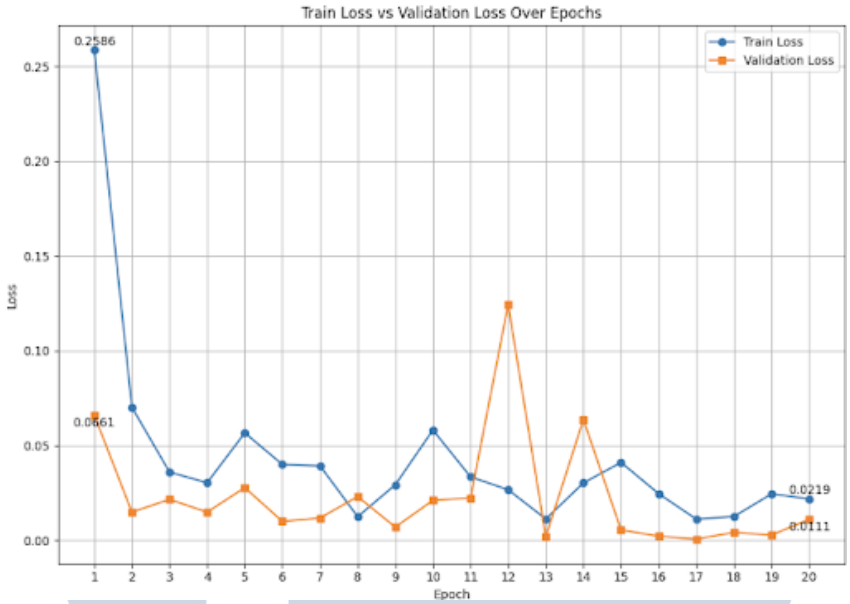
Tabel 3.3. Metrik kinerja selama beberapa epoch

Epoch	Train Loss	Validation Loss	Train Accuracy	Validation Accuracy	Train Precision	Validation Precision	Training Recall	Validation Recall	Training F1	Validation F1
1	0.2586	0.0661	0.9668	0.9914	0.9668	0.9915	0.9668	0.9914	0.9668	0.9914
2	0.0702	0.0149	0.9889	0.9965	0.9889	0.9965	0.9889	0.9965	0.9889	0.9965
3	0.0360	0.0216	0.9947	0.9967	0.9947	0.9967	0.9947	0.9967	0.9947	0.9967
4	0.0304	0.0149	0.9955	0.9990	0.9955	0.9990	0.9955	0.9990	0.9955	0.9990
5	0.0567	0.0278	0.9927	0.9978	0.9927	0.9978	0.9927	0.9978	0.9927	0.9978
6	0.0400	0.0100	0.9965	0.9985	0.9965	0.9985	0.9965	0.9985	0.9965	0.9985
7	0.0393	0.0118	0.9950	0.9993	0.9950	0.9993	0.9950	0.9993	0.9950	0.9993
8	0.0125	0.0231	0.9990	0.9992	0.9990	0.9992	0.9990	0.9992	0.9990	0.9992
9	0.0293	0.0070	0.9977	0.9993	0.9977	0.9993	0.9977	0.9993	0.9977	0.9993
10	0.0582	0.0211	0.9947	0.9978	0.9947	0.9978	0.9947	0.9978	0.9947	0.9978
11	0.0335	0.0224	0.9973	0.9964	0.9973	0.9964	0.9973	0.9964	0.9973	0.9964
12	0.0267	0.1244	0.9983	0.9912	0.9983	0.9913	0.9983	0.9912	0.9983	0.9912
13	0.0113	0.0019	0.9983	0.9993	0.9983	0.9993	0.9983	0.9993	0.9983	0.9993
14	0.0303	0.0637	0.9983	0.9934	0.9983	0.9935	0.9983	0.9934	0.9983	0.9934
15	0.0411	0.0055	0.9965	0.9995	0.9965	0.9995	0.9965	0.9995	0.9965	0.9995
16	0.0246	0.0022	0.9983	0.9993	0.9983	0.9993	0.9983	0.9993	0.9983	0.9993
17	**0.0112**	**0.0007**	**0.9988**	**0.9997**	**0.9988**	**0.9997**	**0.9988**	**0.9997**	**0.9988**	**0.9997**
18	0.0126	0.0042	0.9985	0.9997	0.9985	0.9997	0.9985	0.9997	0.9985	0.9997
19	0.0245	0.0028	0.9973	0.9993	0.9973	0.9993	0.9973	0.9993	0.9973	0.9993
20	0.0219	0.0111	0.9982	0.9983	0.9982	0.9983	0.9982	0.9983	0.9982	0.9983

Tabel 3.3 menyajikan ringkasan lengkap metrik performa model BLOOM-560m pada setiap epoch selama proses penyetelan hiperparameter, memperlihatkan perkembangan performa pelatihan dan validasi. Proses tuning dimulai dengan nilai loss validasi sebesar 0,0661 pada epoch pertama, yang kemudian menunjukkan perbaikan stabil seiring bertambahnya epoch. Pada epoch ke-17, loss validasi mencapai nilai terendah yaitu 0,0007, yang ditetapkan sebagai epoch dengan performa terbaik. Pada titik ini, model berhasil meraih skor nyaris sempurna untuk seluruh metrik validasi, termasuk akurasi (99,97%), presisi (99,97%), recall (99,97%), dan F1-score (99,97%). Hasil ini menunjukkan bahwa model tidak hanya belajar secara efektif dari data pelatihan, tetapi juga mampu melakukan generalisasi dengan sangat baik terhadap data validasi.

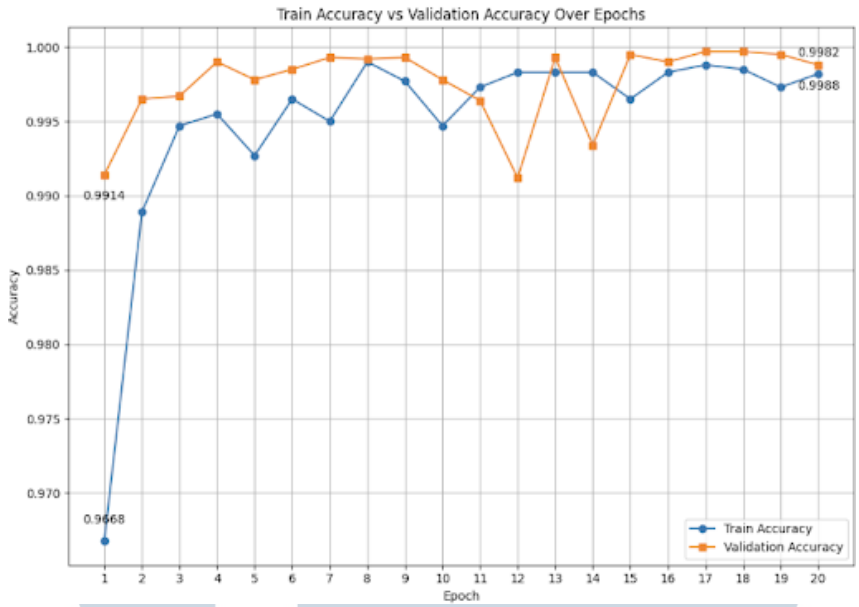
Sepanjang pelatihan, metrik seperti loss pelatihan, akurasi, dan F1-score menunjukkan peningkatan yang konsisten. Loss pelatihan menurun drastis dari 0,2586 pada epoch pertama menjadi 0,0112 pada epoch ke-17. Tren ini menegaskan efektivitas pemilihan hiperparameter, khususnya nilai *learning rate* sebesar `5e-5` dan ukuran *batch* 8, dalam memfasilitasi konvergensi dan kemampuan generalisasi model. Stabilitas akurasi dan presisi validasi yang mulai terlihat sejak sekitar epoch ke-9 mencerminkan kapabilitas pembelajaran yang kuat dari model, sementara penurunan konsisten pada loss validasi memastikan minimisasi kesalahan selama proses pelatihan. Penetapan epoch ke-17 sebagai yang terbaik menunjukkan pentingnya pemantauan metrik loss validasi dan metrik performa lainnya untuk mencapai keseimbangan optimal antara kompleksitas model dan kemampuannya untuk digeneralisasi.

Visualisasi tren metrik pelatihan dan validasi disajikan dalam beberapa grafik berikut, yang menunjukkan perkembangan model selama proses fine-tuning. Setiap grafik memetakan dinamika perubahan metrik utama seperti loss, akurasi, presisi, recall, dan F1-score, baik pada data pelatihan maupun data validasi, selama 20 epoch. Melalui visualisasi ini, dapat diamati pola konvergensi model, stabilitas performa, serta indikasi awal terjadinya overfitting atau underfitting. Pola-pola tersebut memberikan wawasan tambahan terhadap efektivitas proses pelatihan, membantu mengonfirmasi bahwa model telah mencapai titik optimal pada epoch ke-17. Grafik juga membantu membandingkan sejauh mana metrik pelatihan dan validasi saling mendekati, yang merupakan indikator penting dalam menilai kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.



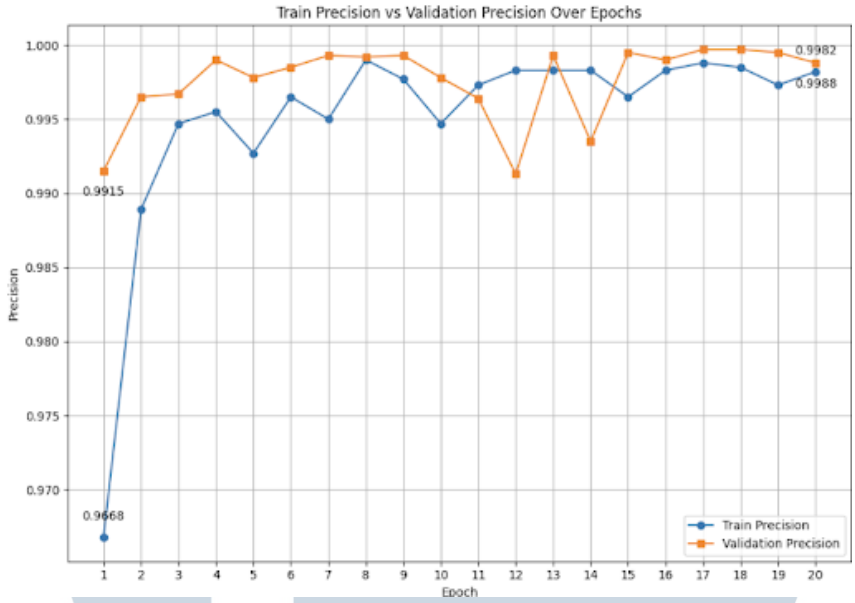
Gambar 3.7. Loss pelatihan dan validasi selama 20 epoch

Gambar 3.7 memperlihatkan perkembangan nilai loss pelatihan dan validasi selama 20 epoch untuk model BLOOM-560m, yang menyoroti tren penting dalam proses optimasi. Loss pelatihan dimulai pada angka 0,2586, sementara loss validasi pada 0,0661. Penurunan signifikan terlihat sejak epoch ke-2. Nilai loss pelatihan terus menurun secara stabil, sedangkan loss validasi sempat mengalami sedikit fluktuasi, dengan kenaikan sementara pada epoch ke-5 dan ke-13. Penurunan tajam terjadi pada loss validasi di sekitar epoch ke-7, mencapai titik terendah sebesar 0,0007 pada epoch ke-17 yang ditetapkan sebagai epoch dengan performa terbaik. Setelah epoch ke-17, loss validasi sedikit meningkat menjadi 0,0110 pada epoch ke-20, sementara loss pelatihan stabil di angka 0,0219. Konsistensi antara kedua nilai loss pada epoch-epoch akhir mencerminkan proses pembelajaran yang efektif dan kemampuan generalisasi model yang kuat.



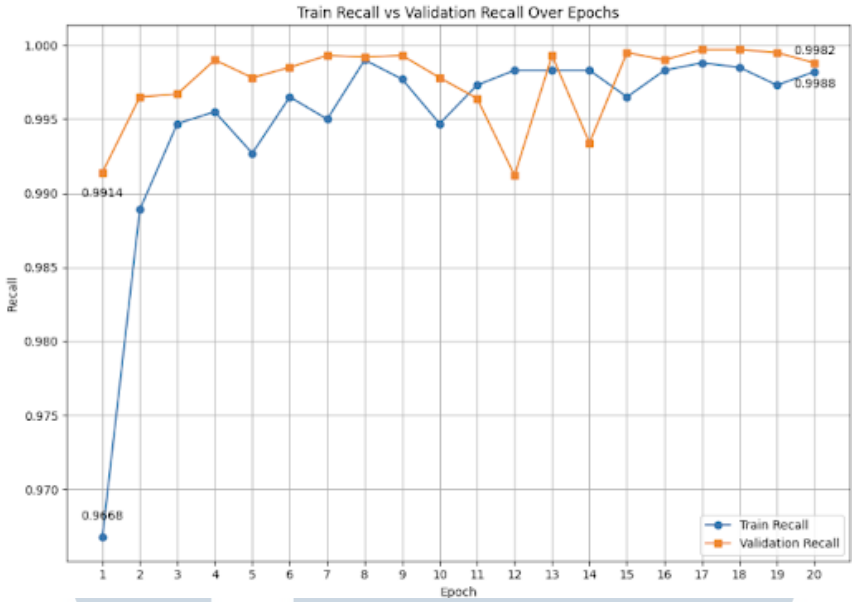
Gambar 3.8. Akurasi pelatihan dan validasi selama 20 epoch

Gambar 3.8 menunjukkan perkembangan akurasi pelatihan dan validasi selama 20 epoch untuk model BLOOM-560m. Akurasi pelatihan dimulai dari 96,68% pada epoch pertama dan meningkat secara konsisten, mencapai puncaknya di 99,88% pada epoch ke-17 dan tetap stabil setelahnya. Akurasi validasi dimulai dari 99,14%, mengalami fluktuasi kecil pada awal epoch dan sempat turun ke 99,00% pada epoch ke-3, kemudian meningkat secara bertahap hingga mencapai nilai tertinggi 99,97% pada epoch ke-17. Mulai dari titik tersebut, kedua metrik stabil, menunjukkan pembelajaran yang efektif dan kemampuan generalisasi yang baik dari model. Fluktuasi awal mencerminkan proses penyesuaian model terhadap dataset, sementara konvergensi akhir menandakan performa yang stabil dan unggul.



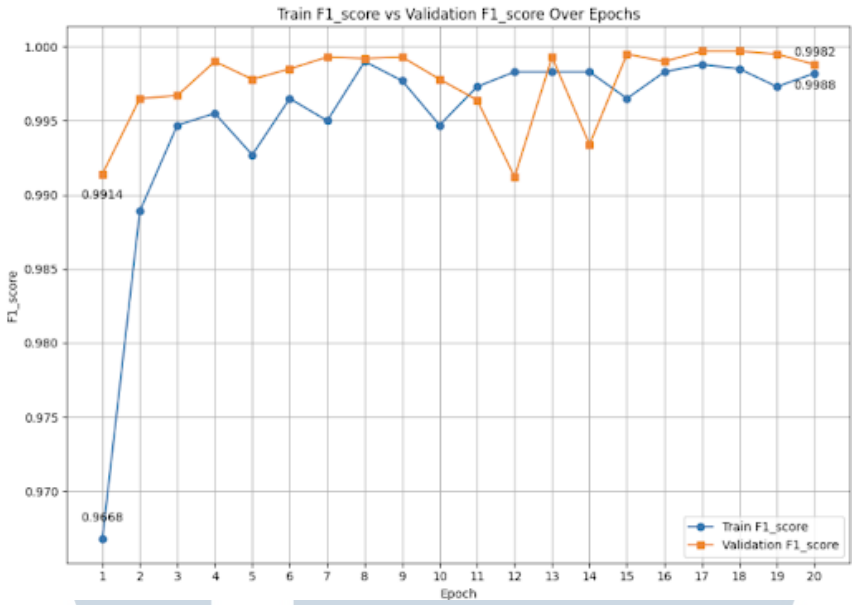
Gambar 3.9. Presisi pelatihan dan validasi selama 20 epoch

Gambar 3.9 menggambarkan perkembangan nilai presisi pelatihan dan validasi selama 20 epoch untuk model BLOOM-560m. Presisi pelatihan dimulai dari 96,68% pada epoch pertama dan meningkat secara bertahap hingga mencapai 99,88% pada epoch ke-17, lalu stabil. Presisi validasi dimulai dari 99,15%, mengalami fluktuasi kecil dengan penurunan ke 99,13% pada epoch ke-12, lalu meningkat hingga mencapai puncaknya sebesar 99,97% pada epoch ke-17. Tren ini menyoroti pembelajaran yang efektif dan kemampuan generalisasi model, dengan fluktuasi awal mencerminkan proses adaptasi terhadap data dan konvergensi akhir yang menunjukkan kinerja presisi yang tinggi dan stabil di kedua set data.



Gambar 3.10. Recall pelatihan dan validasi selama 20 epoch

Gambar 3.10 menampilkan perkembangan nilai recall pelatihan dan validasi selama 20 epoch. Recall pelatihan dimulai dari 96,68% di epoch pertama, meningkat secara stabil hingga mencapai 99,88% pada epoch ke-17 dan tetap stabil. Recall validasi dimulai dari 99,13%, mengalami sedikit fluktuasi awal dengan penurunan ke 99,12% pada epoch ke-12, kemudian meningkat dan mencapai nilai tertinggi 99,96% pada epoch ke-17. Konvergensi antara nilai recall pelatihan dan validasi pada epoch-epoch akhir menunjukkan kemampuan generalisasi yang kuat dari model, dengan fluktuasi awal yang wajar dalam proses penyesuaian terhadap data.



Gambar 3.11. F1-Score pelatihan dan validasi selama 20 epoch

Gambar 3.11 menunjukkan perkembangan skor F1 pelatihan dan validasi selama 20 epoch. Skor F1 pelatihan dimulai dari 96,68% pada epoch pertama, meningkat secara bertahap hingga mencapai 99,88% pada epoch ke-17 dan stabil. Skor F1 validasi dimulai dari 99,13%, mengalami sedikit fluktuasi selama epoch awal dengan penurunan ke 99,12% di epoch ke-12, dan kemudian meningkat hingga puncaknya 99,96% di epoch ke-17. Tren ini menunjukkan pembelajaran yang efektif dan kemampuan generalisasi model, dengan keselarasan skor F1 di kedua set data pada akhir epoch yang mencerminkan kinerja yang konsisten dan kuat.

3.5 Sumber Daya Komputasi

Proses fine-tuning dan penyetelan hiperparameter dalam penelitian ini dilakukan menggunakan layanan Google Colab Pro, dengan memanfaatkan infrastruktur berbasis cloud yang menyediakan akses ke GPU berkinerja tinggi. Konfigurasi spesifik dan penggunaan sumber daya untuk masing-masing proses dijelaskan secara rinci sebagai berikut.

3.5.1 Fine-Tuning

Proses fine-tuning model BLOOM-560M dilakukan dengan parameter *batch size* sebesar 16 dan *learning rate* sebesar 5e-5. Eksperimen ini dijalankan

menggunakan Google Colab Pro dengan konfigurasi perangkat keras sebagai berikut:

- **GPU:** NVIDIA A100 dengan kapasitas VRAM sebesar 40 GB.
- **RAM Sistem:** 83,5 GB tersedia, dengan penggunaan sekitar 4–5 GB selama proses pelatihan.
- **Penggunaan VRAM:** Proses fine-tuning mengonsumsi sekitar 22 GB VRAM, memberikan ruang sisa yang cukup untuk proses komputasi GPU lainnya.
- **Dataset:** 21.000 data untuk pelatihan dan 6.000 data untuk validasi.
- **Waktu Pelatihan:** Setiap epoch memerlukan waktu sekitar 11 menit, dengan total 11 epoch diselesaikan dalam waktu kurang lebih 120 menit.

3.5.2 Penyetelan Hiperparameter

Proses penyetelan hiperparameter dilakukan dengan mengoptimalkan 10 konfigurasi yang berbeda dengan variasi parameter sebagai berikut:

- **Batch Size:** [8, 16]
- **Learning Rate:** Rentang logaritmik dari 5e-1 hingga 5e-5

Karena kebutuhan VRAM yang tinggi (melebihi 40 GB) saat mencoba menjalankan beberapa konfigurasi secara bersamaan, setiap kombinasi parameter dijalankan secara berurutan (sekuensial). Penggunaan sumber daya komputasi untuk setiap skenario dirangkum sebagai berikut:

- **Batch Size 8:** Proses pelatihan dengan batch size 8 memerlukan waktu sekitar 7–8 menit per epoch. Konfigurasi ini tergolong ringan, mengonsumsi sekitar 30 GB VRAM pada GPU NVIDIA A100 dan sekitar 4–5 GB RAM sistem.
- **Batch Size 16:** Saat menggunakan batch size 16, konsumsi VRAM meningkat secara signifikan, mendekati batas maksimum 40 GB dari GPU A100. Hal ini menyebabkan waktu pelatihan menjadi lebih lama, yaitu sekitar 13–15 menit per epoch.

Setiap konfigurasi dijalankan selama beberapa epoch untuk mengevaluasi performa model dan mengoptimalkan parameter. Rata-rata, dua epoch per konfigurasi cukup untuk mengamati pola konvergensi, sehingga total waktu penyetelan hiperparameter per konfigurasi berkisar antara 2 hingga 3 jam.