LAPORAN FINE-TUNING LLM

NUR ALMAR'ATUSSALIHA

23917040

MAGISTER INFORMATIKA

2025

1. Pendahuluan

Pada pemrosesan bahasa alami (Natural Language Processing), pengelompokan kata berdasarkan fungsi tata Bahasa dikenal sebagai Part of Speech (POS) Tagging, merupakan langkah penting untuk memahami konteks suatu kalimat [1]. POS Tagging berperan dalam menentukan kategori tata bahasa dari kata-kata dalam teks, seperti kata benda (*noun*), kata kerja (*verb*), kata sifat (*adjective*), dan sebagainya [2]. Seiring dengan perkeembangan teknologi, pendekatan pembelajaran berbasis mendalam (*Deep Learning*), seperti *model Bidirectional Encoder Representations from Transformers* (BERT) dan *Robustly Optimized BERT Approach* (RoBERTa), telah menunjukkan kinerja yang sangat baik dalam tugas-tugas NLP, termasuk POS Tagging. Model ini menggunakan arsitektur transformer untuk menangkap hubungan semantik dan sintaksis antar kata secara kontekstual [3][4].

Dalam beberapa tahun terakhir, teknik fine-tuning telah menjadi metode yang efektif untuk memanfaatkan kekuatan model bahasa besar (*Large Language Models*). Dengan memanfaatkan *pre-trained* models seperti BERT dan RoBERTa, penyesuaian model pada dataset khusus memungkinkan peningkatan performa seperti akurasi dan efisiensi dalam tugas POS Tagging [5]. Selain itu, teknik seperti *Parameter-Efficient Fine-Tuning* (PEFT) menawarkan pendekatan yang lebih hemat sumber daya. Salah satu metode PEFT yang menonjol adalah LoRA (*Low-Rank Adaptation*), yang memungkinkan *fine-tuning* model besar dengan memperkenalkan adaptor parameter rendah ke dalam arsitektur model utama. LoRA memberikan solusi yang lebih ringan dan efisien untuk tugas-tugas NLP, termasuk POS Tagging, terutama dalam skenario dengan keterbatasan perangkat keras [6].

Penulisan laporan ini bertujuan untuk membandingkan kinerja model BERT dan RoBERTa setelah diterapkan teknik fine-tuning pada dataset POS Tagging. Selain itu, laporan ini juga mengevaluasi performa Parameter-Efficient Fine-Tuning (PEFT) dengan metode *Low-Rank Adaptation* (LoRA) terhadap efisiensi dan akurasi kedua model tersebut. Dengan membandingkan hasil *fine-tuning* konvensional dan PEFT, hasil dari laporan ini diharapkan utuk memberi memberikan wawasan tentang pendekatan yang efektif dan efisien untuk tugas POS Tagging dengan memanfaatkan sumber daya terbatas.

2. Model

2.1.BERT

Bidirectional Encoder Representation from Transformers (BERT) adalah model Large Language Model (LLM) berbasis arsitektur Transformrt yang dikembangkan oleh Jacob Devlin dan timnya. BERT dirancang untuk memahami konteks kata dalam kalimat secara *bidirectional*, yaitu dari kiri ke kanan dan sebaliknya, sehingga menghasilkan pemahaman kontekstual lebih baik dibandingkan model sebelumnya [7].

Pre-trained BERT dilakukan pada kumpulan data teks yang sangat besar atau sering disebut korpus untuk menangkap pola semantic dant sintaksis umum dengan banyak tugas NLP [7]. Model ini dapat disesuaikan untuk tugas spesifik menggunakan *fine-tuning*, proses yang membutuhkan dataset lebih kecil dibandingkan dengan *training* dari awal [8]. BERT juga dapat diadaptasi untuk berbagai tugas seperti klasifikasi teks dan POS tagging [9].

2.2.RoBERTa

Robustly Optimized BERT Pretraining Approach (RoBERTa) adalah model berbasis transformer yang dikembangkan oleh Liu dan timnya sebagai penyempurna model BERT [10]. RoBERTa dilatih pada Kumpulan data yang jauh lebih besar dibandingkan BERT dengan total data sekitar 160 GB sehingga waktu *training* jauh lebih Panjang agar dapat memaksimalkan pembelajaran dari data [10]. Berbeda dengan BERT yang menggunakan *static masking*, RoBERTa menerapkan *dynamic maskin*g di mana pola *masking* berubah setiap kali pelatihan diproses sehingga meningkatkan keragaman data yang dilihat model [10].

Dalam studi pada Bahasa dengan sumber daya terbatas, RoBERTa memiliki kinerja yang lebih baik karena memanfaatkan transfer learning dari data yang lebih besar [11]. RoBERTa sering kali mengungguli BERT dalam tugas POS Tagging karena kemampuan dalam penghapusan NSP dan penggunaaan dataset yang lebih besar[12]

2.3. LoRA

Low-rank Adaptation (LoRA) adalah Teknik Parameter-Efficient Fine-Tuning (PEFT) untuk mengurangi biaya komputasi saat fine-tuning model LLM. LoRA menyesuaikan model dengan menambahkan adaptor berbobot rendah ke lapisan inti model tanpa memperbaharui semua parameter inti [13]. Model BERT dan RoBERTa memiliki ratusan juta parameter, fine-tuning penuh membutuhkan memori besar dan waktu pelatihan yang lama. LoRA dapat menyesuaikan model hanya dengan melatih adaptop kecil sehingga mengurangi beban komputasi [14].

Penerapan LoRA pada model BERT untuk tugas POS tagging dapat menunjukkan kinerja yang sama dengan menggunakan fine-tuning penuh namun dengan penggunaan komputasi yang lebih rendah [14]. Penggunaan LoRA pada model RoBERTa juga dapat menunjukkan peningkatan akurasi yang signifikan, sama halnya jika melakukan fine-tuning penuh [15].

3. Eksperimen

Ekperimen fine-tuning POS tagging menggunakan dataset yang sudah ada dari huggingface, Data tersebut terdiri data *training* sebanyak 13.100 baris dan data *testing* sebanyak 1450 baris. Terdapat 2 feature yaitu words yang berisi kata yang telah berbentuk token, *feature* label yang berisi kata yang dianotasi dengan tag. Berikut contoh dari dataset tersebut:

Tabel 1: Contoh dataset

Words	Labels
["Confidence", "in", "the", "pound", "is", "widely", "expected", "to", "take", "another", "sharp", "dive", "if", "trade", "figures", "for", "September", ",", "due", "for", "release", "tomorrow", ",", "fail", "to", "show", "a", "substantial", "improvement", "from", "July",	["NN", "IN", "DT", "NN", "VBZ", "RB", "VBN", "TO", "VB", "DT", "JJ", "NN", "IN", "NN", "NNS", "IN", "NNP", ",", "JJ", "IN", "NN", "NN", ",", "VB", "TO", "VB", "DT", "JJ", "NN", "IN", "NNP", "CC", "NNP", "POS", "JJ", "NNS", "."]

"and", "August", "'s", "near-record", "deficits", "."]	

Tahap preprocessing yang dilakukan adalah mengecek jumlah token dan tag sama, tokenisasi berbasis subword, melakukan padding atau menyamakan Panjang input, dan *encoding* data yaitu *feature* diubah ke representasi numerik.

Dalam eksperimen ini, skor F1 dipilih untuk mengukur performa karena dianggap lebih cocok dibandingkan dengan akurasi, terutama pada dataset dengan kelas yang tidak seimbang [16].

Tabel 2: Label dan jumlah token

Label	Jumlah Token
NN	44657
IN	33979
NNP	30882
DT	27541
JJ	19706
CD	12055
VBD	10302
RB	9807
VB	8848
CC	8022
TO	7591
VBN	7178
VBZ	6941
PRP	5829
VBG	4928
VBP	4273
MD	3183
PRP$	2811
POS	2679

Tahap modeling menggunakan *transformer* dengan model *pre-trained* BERT dan RoBERTa dari huggingface yaitu; *bert-based case*, dan *roberta base* yang dijalankan melalui Google Colab dan T4 GPU. Adapun hyperparameter yang digunakan yaitu:

- Learning rate: 5e-5
- Batch size: 16
- Epoch: 3
- Weight_decay: 0.01

Adapun hasil *training* model sebagai berikut:

Tabel 3. Hasil fine-tuning model BERT

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	0.084100	0.103885	0.961537	0.961476	0.961506	0.972310
2	0.049600	0.101729	0.964589	0.964929	0.964759	0.974423
3	0.026700	0.107317	0.964212	0.965568	0.964889	0.974384

Tabel 4. Hasil fine-tuning model BERT dengan PEFT LoRA

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	0.023900	0. 103123	0.965050	0.965760	0.965405	0.974897
2	0.022500	0. 104071	0.964588	0.965760	0.965173	0.974858
3	0.024400	0. 104389	0.964021	0.965408	0.964714	0.974660

Dengan meggunakan estimasi parameter LoRA, jumlah parameter yang dilatih (di-update selama *training*) sebanyak 333.824 dengan jumlah keseluruhan parameter 108.088.416, artinya hanya sekitar 0.3070% parameter yang digunakan dan hasilnya cukup optimal, yaitu performa hampir sama dengan *fine-tuning* biasa.

Tabel 5. Hasil Fine-tuning model RoBERTa

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	0.086900	0.096571	0.962745	0.962499	0.962622	0.973558
2	0.049600	0.107621	0.961876	0.962307	0.962092	0.973104
3	0.037300	0.107279	0.964171	0.963586	0.963879	0.973888

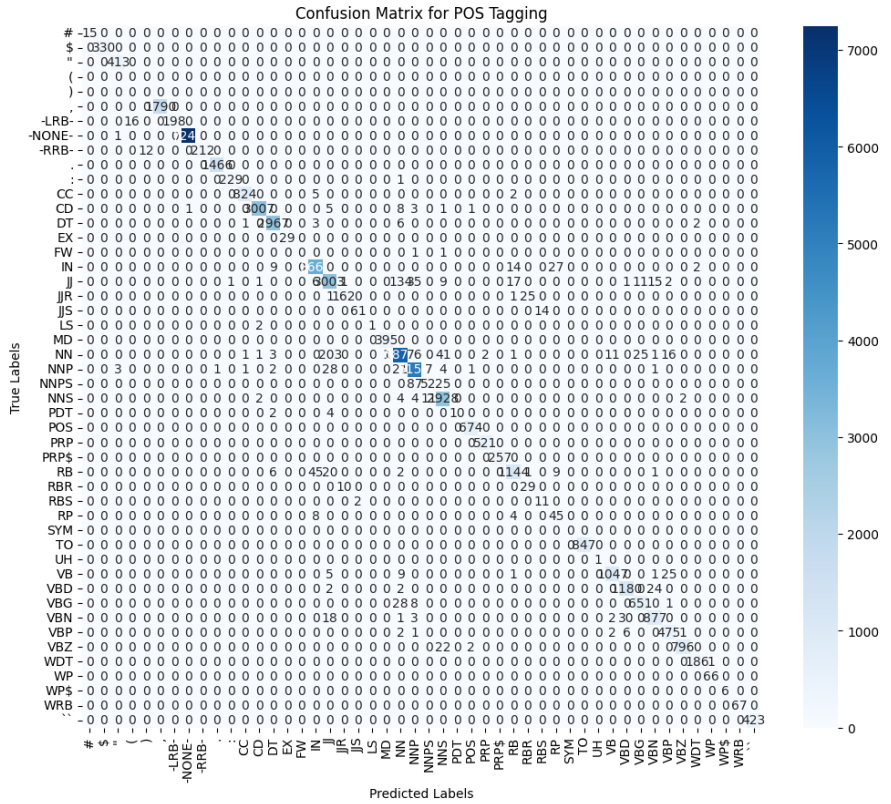
Tabel 6. Hasil Fine-tuning model RoBERTa dengan PEFT LoRA

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	0. 027500	0. 111149	0.964443	0.963426	0.963934	0.973599
2	0. 028400	0. 111944	0.963925	0.963586	0.963755	0.973764
3	0. 032900	0. 112124	0.963863	0.963586	0.963724	0.973743

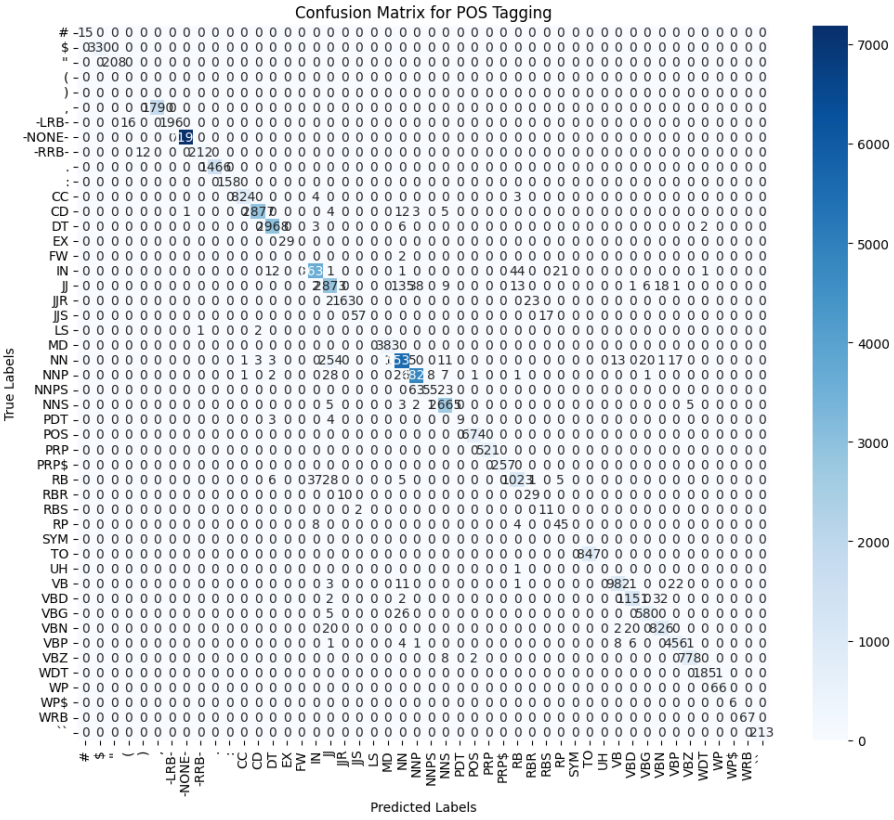
Dengan menggunakan estimasi parameter model LoRA, jumlah parameter yang digunakan pada model RoBERTa selama pelatihan adalah 331.824 dengan jumlah keseluruhan parameter 124.423.776, artinya hanya sekitar 0.2667% parameter yang digunakan selama pelatihan, Hasil tersebut cukup baik jika dibandingkan dengan *full fine-tuning*.

Metrik evaluasi utama (Precision, Recall, F1-Score, dan Accuracy) dari *fine-tuning* dengan LoRA sangat kompetitif dibandingkan dengan *fine-tuning* biasa, bahkan dengan selisih yang sangat kecil (hampir tidak signifikan). Pada model BERT dan RoBERTa, LoRA mampu mencapai tingkat akurasi yang hampir sama atau sedikit lebih rendah (perbedaan dalam kisaran 0.001–0.002), namun tetap berada dalam *range* yang dapat diterima.

Fine-tuning dengan LoRA menunjukkan stabilitas yang lebih baik, terutama dalam hal *training loss* dan *validation loss*, yang tidak meningkat secara signifikan pada *epoch* berikutnya. *Training loss* lebih rendah dengan LoRA, menunjukkan model belajar lebih cepat dan efektif.



Gambar 1. Confusion matrix model BERT



Gambar 2. Confusion matrix model RoBERTa

4. Kesimpulan

Fine-tuning biasa dan *fine-tuning* dengan LoRA (*Low-Rank Adaptation*) sama-sama memberikan hasil yang sangat baik. LoRA memberikan hasil yang lebih konsisten selama pelatihan, terutama pada nilai *training loss* dan *validation loss*, yang tidak terlalu fluktuatif. Ini menunjukkan bahwa LoRA lebih bisa menghindari model dari *overfitting*.

Menggunakan LoRA untuk *fine-tuning* adalah pilihan yang sama baiknya, atau bahkan lebih baik daripada fine-tuning biasa. LoRA sangat cocok untuk tugas-tugas yang membutuhkan efisiensi dan stabilitas, tanpa mengorbankan kualitas hasil.

Referensi

[1] Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing. Pearson Education.

[2] Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). "Feature-rich part-of-speech tagging with a cyclic dependency network." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT 2019*, 4171–4186.

[4] 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*.

[5] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). "How to Fine-Tune BERT for Text Classification?" *China National Conference on Chinese Computational Linguistics*. Springer.

[6] Hu, E. J., Shen, D., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., & Chen, W. (2021). "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv preprint arXiv:2106.09685*.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT 2019*.

[8] Howard, J., & Gugger, S. (2018). "Universal Language Model Fine-tuning for Text Classification." *Proceedings of ACL 2018*.

[9] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach."

[10] 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*.

[11] Wu, S., Dredze, M., & Zhang, J. (2020). "Multilingual BERT Fine-tuning for Low-Resource Languages." *Proceedings of ACL 2020.*

[12] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). "How to Fine-Tune BERT for Text Classification?" *China National Conference on Chinese Computational Linguistics.*

[13] Hu, E. J., Shen, D., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., & Chen, W. (2021). "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv preprint arXiv:2106.09685.*

[14] Ding, J., Chen, W., Wang, H., & Chen, L. (2022). "Parameter-Efficient Fine-Tuning Using Low-Rank Adaptation for Multi-Domain Language Models." *Proceedings of NAACL-HLT 2022.*

[15] Wu, S., Dredze, M., & Zhang, J. (2020). "Multilingual BERT Fine-tuning for Low-Resource Languages." *Proceedings of ACL 2020.*

[16] Seok, M., Song, H.-J., Park, C.-Y., Kim, J.-D., & Kim, Y.-s. (2016). Named entity recognition using word embedding as a feature. *International Journal of Software Engineering and Its Applications, 10*(2), 93–104.