Computer Science and Information Technologies
Vol. 6, No. 3, November 2025, pp. 274~282
ISSN: 2722-3221, DOI: 10.11591/csit.v6i3.p274-282

Optimizing diplomatic indexing: full-parameter vs low-rank adaptation for multi-label classification of diplomatic cables

Dela Nurlaila, Abba Suganda Girsang
Department of Computer Science, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Sep 19, 2024
Revised Jun 9, 2025
Accepted Jun 13, 2025

Keywords:

Diplomatic cables
Diplomatic index
Full-parameter tuning
Low-rank adaptation
Multi-label classification

ABSTRACT

Accurate classification of diplomatic cables is crucial for Mission's evaluation and policy formulation. However, these documents often cover multiple topics, hence a multi-label classification approach is necessary. This research explores the application of pre-trained language models (CahyaBERT, IndoBERT, and MBERT) for multi-label classification of diplomatic cable executive summaries, which align with the diplomatic representation index. The study compares full-parameter fine-tuning and low-rank adaptation (LoRA) techniques using cables from 2022-2023. Results demonstrate that Indonesian-specific models, particularly the IndoBERT, outperform multilingual models in classification accuracy. While LoRA showed slightly lower performance than full fine-tuning, it significantly reduced GPU memory usage by 48% and training time by 69.7%. These findings highlight LoRA's potential for resource-constrained diplomatic institutions, advancing natural language processing in diplomacy and offering pathways for efficient, real-time multi-label classification to enhance diplomatic mission evaluation.

Corresponding Author:

Dela Nurlaila
Department of Computer Science, BINUS Graduate Program-Master of Computer Science
Bina Nusantara University
Jakarta 11480, Indonesia
Email: dela.nurlaila@binus.ac.id

1. INTRODUCTION

Freedom of communication is guaranteed to diplomatic missions by Article 27 of the Vienna Convention of 1961 [1]. How diplomacy works has changed since technological advancements existed [2]. Nevertheless, diplomatic cables remain an important form of communication in the world of diplomacy, even though the communication technology being used has evolved. A cable represents diplomatic correspondence between a foreign ministry in the home country and its diplomatic missions abroad [3]. Through these cables, sensitive information is exchanged, an in-depth analysis of the accredited countries is reported, potential cooperation is explored, and instructions are given from the home government. The vast amount of data exchanged provides a rich source of information for evaluating diplomatic missions. According to Bjola *et al.* [4], in today's era, data is "the new oil", and a significant asset to its owner. The information gleaned from diplomatic cables can be used to gauge a Mission's performance. The diplomatic representation index serves as a numerical assessment system that covers a range of performance indicators for areas including politics, economy, protocol and consular services, socio-cultural affairs, and administration.

However, accurately classifying and indexing this wealth of information poses a significant challenge. The complexity of diplomatic work, spanning multiple functional domains, makes manual classification inefficient and prone to error. Accurate classification is essential not only to support the

diplomatic representation index, which is used to evaluate the performance of diplomatic missions in different areas but also to ensure that foreign policy decisions are based on the right information. On the other hand, misclassification could result in incorrect evaluations of a mission's success, misguided policy decisions, and even pose risks to national security.

Multi-label classification is a technique in machine learning that enables assigning multiple categories to a single instance [5]. Using this supervised learning approach [6], diplomatic documents can be categorized into multiple topics that represent the various natures of diplomatic work. This research explores two fine-tuning approaches in the context of multi-label classification of diplomatic cables: full parameter fine-tuning [7], [8], and the more efficient low-rank adaptation (LoRA) [9] method. When applied to the BERT model [10], which is known for its robust performance on a variety of natural language processing (NLP) tasks, both methods have shown promising results. Fine-tuning is a process of adjusting a pre-trained model's parameters to improve its performance on a specific task, while, LoRA updates only specific parameters, allowing the model to quickly adapt to specific tasks without requiring extensive computational resources while achieving competitive performance results compared to full parameter fine-tuning method.

Multi-label text classification of Indonesian customer reviews using IndoBERT as an end-to-end model improved with an accuracy of up to 19.19%, compared to using IndoBERT with CNN and XGBoost classifier [11]. In Nabiilah *et al*. [12], multi-label classification of toxic comments in Indonesian social media by using IndoBERT for feature extraction and MBERT for classification achieved optimal results with an F1 score of 0.9032. In the context of Indonesian language processing, pre-trained models such as CahyaBERT have demonstrated promising performance across various NLP tasks, such as for name entity recognition on hadith texts achieved impressive F1-scores of 99.63% in testing [13]. While focused on NER, this success has inspired the application of CahyaBERT for multi-label classification tasks. LoRA is used in [14] for classifying legal documents. The results show that LoRA performs better than full parameter fine-tuning while requiring fewer computational resources, with a rank of 32. The researcher developed a dataset called Taiwan Legal Judgement Prediction (TWLJP) and compared the performance of LoRA against full parameter fine-tuning methods and other models like Lawformer. Results show that LoRA achieved comparable performance to full fine-tuning while significantly reducing computational resources and training time, requiring only about 0.72% of the trainable parameters compared to the full model.

This article will compare cahya/bert-base/indonesian-1.5G, indolem/indobert-base-uncased, and bert-base-multilingual-cased using full parameter fine-tuning and LoRA to classify the diplomatic cables, into multi-labels categories, which is politics, economics, protocol and consular services, socio-cultural affairs, and administrative matters. The dataset used is diplomatic cables sampled from 2022-2023, and only the executive summaries sections are taken. This study will be conducted using a single NVIDIA GeForce GTX 1080 8GB GPU to enable comprehensive evaluation trade-offs between model performance and resource utilization, which is crucial for diplomatic institutions that have limited computational resources. Accurately determining and assessing the categories of diplomatic cables is essential so that diplomatic activities can align with the national foreign policy and contribute to international relations strategy.

2. METHOD

Figure 1 shows the steps for this research in general. In data collection, the executive summaries are extracted from the document samples and annotated. The next step is data preparation, which includes pre-processing, data splitting, and exploratory data analysis (EDA). Two different modeling techniques are applied to the model, full-parameter fine-tuning and fine-tuning with LoRA. The final step is evaluating both models.

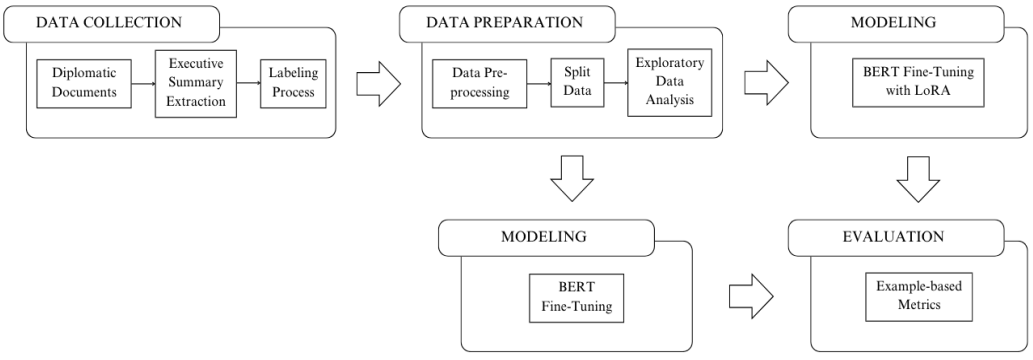


Figure 1. Steps in the proposed method

2.1. Data collection

The dataset utilized for this study was obtained from the Ministry of Foreign Affairs of the Republic of Indonesia, comprising 1,329 diplomatic cables collected from 130 Missions. The executive summaries of all documents were extracted and categorized using a multi-labeling system. Table 1 displays samples of labeled data. Each sample may be associated with multiple categories that reflect the overlapping themes present in the executive summaries.

Table 1. Example of data labeling

No	Summary	Label
B-00327/Frankfurt/231124	The Indonesian Consulate General in Frankfurt held a virtual discussion with BNI London entitled "Developing Indonesian Diaspora Entrepreneurship in the Consulate General's Area of Responsibility" on November 23, 2023. The event was opened by the Indonesian Consul General in Frankfurt and featured speakers from BNI London and BNI Headquarters.....	economic, socio-cultural
B-00268/Bratislava/231227	The Indonesian Embassy in Bratislava held a community outreach event in the form of a Christmas service and celebration on December 19, 2023. The event was organized in collaboration with the Indonesian Christian Family in Slovakia (Garisindo). The event was attended by around 100 Indonesians and members of the Indonesian diaspora in....	socio-cultural

The dataset encompasses five categories of diplomatic functions: political, economic, socio-cultural, protocol and consular, and administration. Each of these functions is specifically detailed in the regulation [15]. Figure 2 illustrates the distribution of documents based on how many labels each document has, ranging from one to five labels per document. The data reveals that single-label documents are the most prevalent, totaling 842 documents. Documents with dual labels follow, accounting for 323 documents. The number of documents with three labels amounted to 61 documents, with four labels reaching 78, and the least common are documents with five labels, only 25 documents.

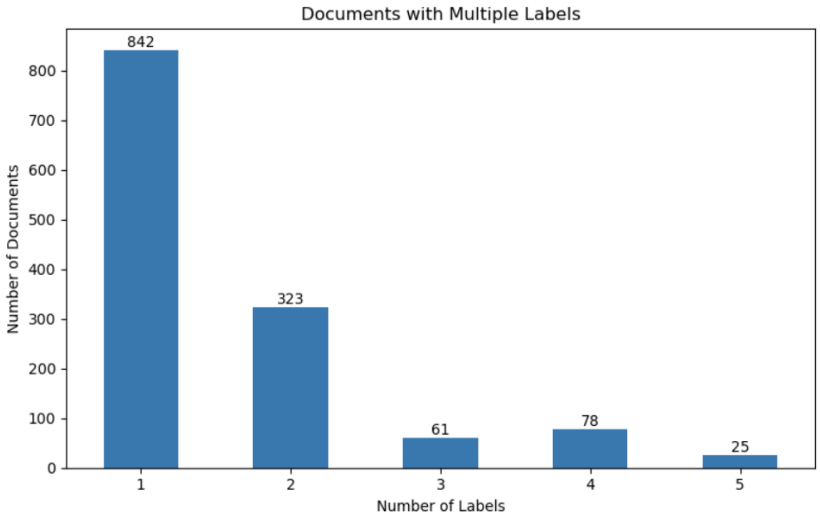


Figure 2. Distribution of the number of labels in the dataset

2.2. Data preparation

In order to preserve its context [16], text pre-processing is minimized, with only removal of URLs from the executive summary. Data splitting is a process of dividing a dataset into separate subsets to reduce bias, prevent overfitting, and accurately evaluate the model performance on unseen data [17]. For this research, a stratified sampling method is applied, where the population of the data is divided into different subgroups and ensures that the final samples reflect the proportions of these subgroups [18]. This approach is crucial for multi-label datasets, as it helps maintain the balance of label combinations across all subsets, ensuring that each subset is representative of the overall data distribution. By applying this approach, the

dataset is split into three subsets: 80% for the training set and the remaining 20% for the testing set, whilst one-tenth of the training set is randomly selected for validation data [19]. However, due to the constraints of maintaining label balance across subsets in multi-label datasets led to a modified distribution, as shown in Table 2. Furthermore, additional EDA was performed to analyze the training data, in order to get more info about the data pattern.

Table 2. Dataset statistic

Multi-label	Dataset	Data train	Data test	Data validation
1 label (single)	842	605	169	68
2 labels	323	232	65	26
3 labels	61	44	12	5
4 labels	78	56	16	6
5 labels	25	18	5	2

2.3. Model settings

In this research, three different NLP models were selected: two BERT-base models trained in monolingual, namely IndoBERT [20] and CahyaBERT [13], as well as one multilingual model, MBERT [21]. These models share the same configuration for embedding size, number of hidden layers, and attention heads which are 768, 12, and 12, respectively. Table 3 provides a summary of the hyperparameter configurations used in prior research to develop the pre-trained BERT that are utilized in this research.

Table 3. Hyperparameter summary of BERT models

Model	Type	Parameters	Corpus size	Vocab size
cahya/bert-base/indonesian-1.5G	Monolingual (Indonesian)	110 M	1.5 G	32,000
indolem/indobert-base-uncased	Monolingual (Indonesian)	110 M	220 M	31,923
bert-base-multilingual-cased	Multilingual	172 M	Wikipedia (104 languages)	119,547

Two prominent approaches were used for this study, which are full-parameter fine-tuning and parameter-efficient tuning, with a particular focus on LoRA. Full-parameter fine-tuning involves updating all parameters of a pre-trained model during the training process, allowing the model to adapt fully to the task [22], [23]. On the other hand, LoRA aims to adapt models with minimal changes to the original parameters, which results in high parameter efficiency [24]. It introduces trainable rank decomposition matrices to the weights, allowing adaptations while keeping most of the original model frozen [9]. Table 4 outlines the hyperparameters used during the fine-tuning process, and Table 5 detailing the hyperparameters specific to LoRA-based tuning. All experiments were conducted using an NVIDIA GeForce GTX 1080 8GB GPU.

Table 4. Hyperparameter settings for full-parameter fine-tuning

Hyperparameter	Value
Learning rate	1e-5
Epoch	30
Batch size	16
Warmup ratio	0.1
LR scheduler type	Linear
Dropout rate	0.3
Weight decay	0.01
Patience	3

Table 5. Hyperparameter settings for LoRA

Hyperparameter	Value
Learning rate	3e-5
Epoch	30
Batch size	4
Warmup ratio	0.1
LR scheduler type	Cosine
Dropout rate	0.3
Weight decay	0.01
Patience	3
Rank	16
Alpha	32
Target modules	Q, K, V

2.4. Evaluation method

In this study, the evaluation metrics used to measure the performance of the multi-label classification task are hamming loss and subset accuracy [6]. Hamming loss evaluates whether each label of each sample is predicted correctly. A lower hamming loss value indicates that the model is more accurate in predicting each label. Subset accuracy evaluates whether the entire set of labels for a given sample is predicted exactly correctly. The higher the subset accuracy value, the better the performance of the model.

3. RESULTS AND DISCUSSION

To determine the optimal token length, a tokenizer from each pre-trained model is used to calculate the distribution of token lengths in the dataset. Table 6 shows the distribution of token counts for each model, where the majority of samples have token lengths under 512, except for MBERT, where only 0.6% of the data are outliers. Therefore, 512 is used as the maximum token length. By maximizing, token lengths helps to capture the essential context of the data, enabling a richer understanding of the text.

Table 6. Distribution of token counts

Model	MinToken	MaxToken	% of entries <=512
cahya/bert-base/indonesian-1.5G	13	437	100
indolem/indobert-base-uncased	14	442	100
bert-base-multilingual-cased	16	568	99.4

Additional EDA was performed to analyze the training data, revealing that it remains imbalanced. To address this uneven distribution, two complementary approaches were applied: data augmentation and random undersampling (RUS). Data augmentation is a technique that increases the amount and diversity of data by modifying existing examples or creating a new one [25]. One common data augmentation method is back translation [26]–[28]. In this study, back translation is used where the original Indonesian text is translated to English, and then converted back to Indonesian. This process is carried out using the Translator class from the googletrans library. A ratio threshold is introduced to ensure the augmentation does not disproportionately increase synthetic data, helping to maintain a balanced dataset. To further balance the data, RUS were implemented. RUS is a method for handling imbalanced data by randomly removing samples from the majority class [29]–[31].

As shown in Figure 3, the process of back translation and RUS has significantly altered the distribution of label combinations in the training data. The most notable changes are observed in single and double label combinations. The original data of 955 samples has been expanded to 1,655 samples, an increase of 73.3%. The least frequent data category containing 5 labels also doubled in quantity.

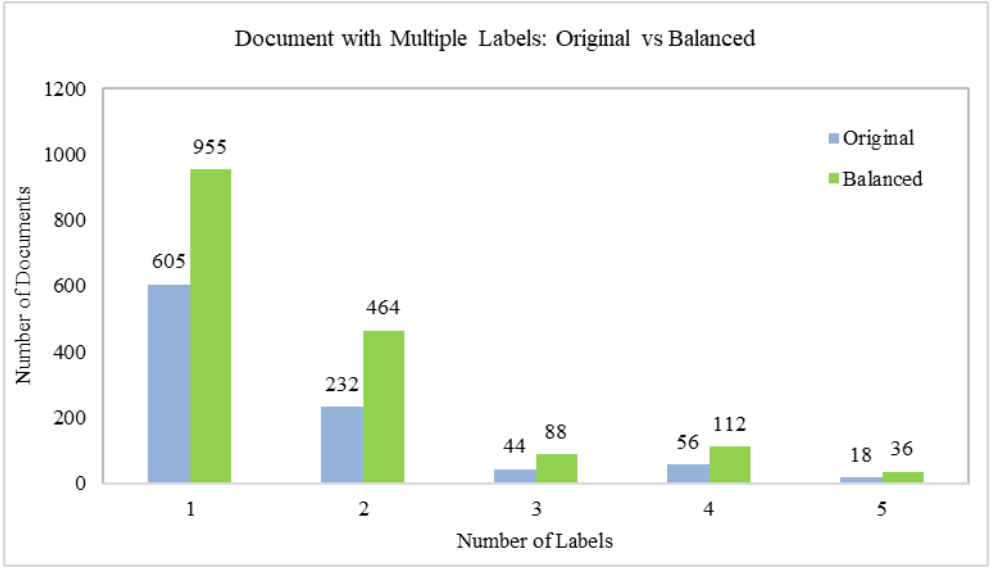
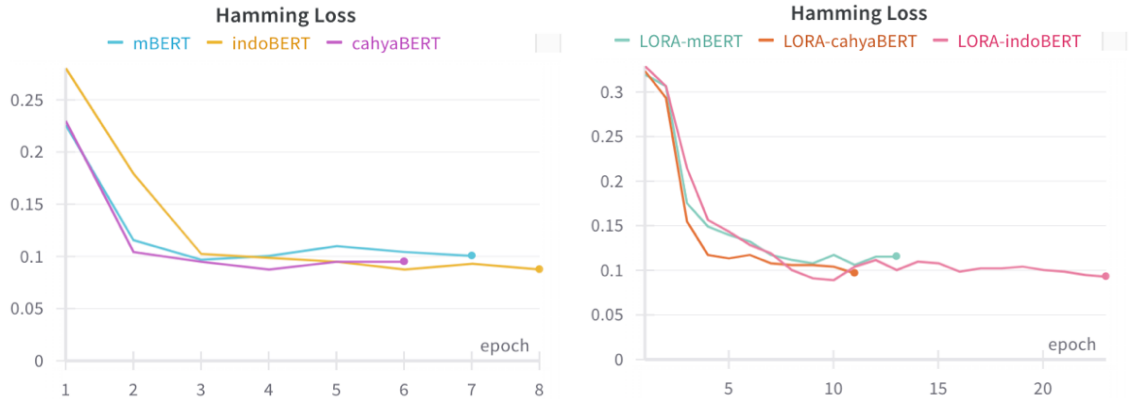


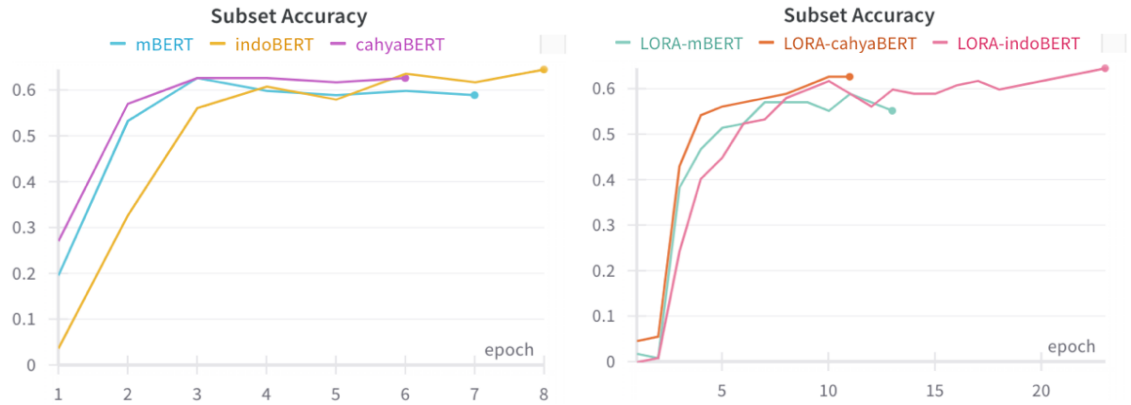
Figure 3. Data training with multiple labels: original vs balanced

Figure 4 shows the hamming loss metrics across various epochs. In the full-parameter fine-tuning, as shown in Figure 4(a), IndoBERT achieves the lowest hamming loss value compared to CahyaBERT and MBERT. While in Figure 4(b), the hamming loss value for LoRA-based tuning demonstrates that LoRA-IndoBERT consistently maintains the lowest loss. Figure 5 compares subset accuracy. The full-parameter fine-tuning in Figure 5(a) IndoBERT starts with the lowest accuracy but surpasses other models by epoch 8. The LoRA-based tuning graph in Figure 5(b) reveals a longer training period, with LoRA-IndoBERT demonstrating the highest final accuracy.



(a) (b)

Figure 4. Hamming loss for (a) full-parameter fine-tuning and (b) LoRA-based tuning



(a) (b)

Figure 5. Subset accuracy for (a) full-parameter fine-tuning and (b) LoRA-based tuning

The experiment as illustrated in Table 7, shows that among the full fine-tuned models, IndoBERT achieves the lowest hamming loss of 0.0831 and shares the same highest value for subset accuracy of 0.6779 with CahyaBERT. This suggests that Indonesian-specific pre-trained models are particularly effective for this task. Interestingly, the implementation of LoRA method shows a slight decrease in the performance of all models. For instance, LoRA-IndoBERT has a higher hamming loss (0.0914) and lower subset accuracy (0.6517) compared to its full fine-tuned model. The MBERT, while competitive, consistently underperforms compared to Indonesian-specific models, highlighting the importance of a language-specific model for this task.

As shown in Table 7, there is a slight performance gap between LoRA and the full fine-tuned model. This difference can be attributed to architectural differences in parameter updating, where full fine-tuning modifies all components of the models, while LoRA only updates parameters in the self-attention layer through LoRA. As reported in Table 8, LoRA only trains about 0.78% of total parameters in CahyaBERT and indoBERT, and 0.49% in MBERT. In the multi-label context, where the data often contain overlapping topics, LoRA's limited parameter space particularly affects its ability to capture complex topic interdependencies between 5 different labels. This architectural constraint becomes evident in our results

where LoRA requires more epochs to achieve optimal performance compared to full fine-tuning, as a trade-off between parameter efficiency and model adaptability.

Table 7. Evaluation results on testing data

Model	Hamming loss	Subset accuracy
Cahya/bert-base/indonesian-1.5G	0.0869	0.6779
Indolem/indobert-base-uncased	0.0831	0.6779
Bert-base-multilingual-cased	0.0974	0.6517
Cahya/bert-base/indonesian-1.5G + lora	0.1019	0.6255
Indolem/indobert-base-uncased + lora	0.0914	0.6517
Bert-base-multilingual-cased + lora	0.1004	0.6217

Table 8. Trainable parameters

Model	Total parameter	LoRA trainable parameter	Trainable (%)
Cahya/bert-base/indonesian-1.5G	113.873.674	888.581	0.7803
Indolem/indobert-base-uncased	113.814.538	888.581	0.7807
Bert-base-multilingual-cased	181.109.770	888.581	0.4906

While LoRA achieves slightly lower performance metrics, Figure 6 shows that LoRA-IndoBERT only uses 52% of GPU memory and training duration completed in under 45 minutes, which is a 69.7% reduction in time compared to full fine-tuning. This efficiency makes LoRA an attractive option for institutions with limited computational resources, as it can effectively run on consumer-grade GPUs. Future improvements could focus on optimizing LoRA's architecture specifically for multi-label classification tasks, such as developing mechanisms to better capture label relationships or exploring selective parameter adaptation in other layers while preserving its core efficiency benefits.

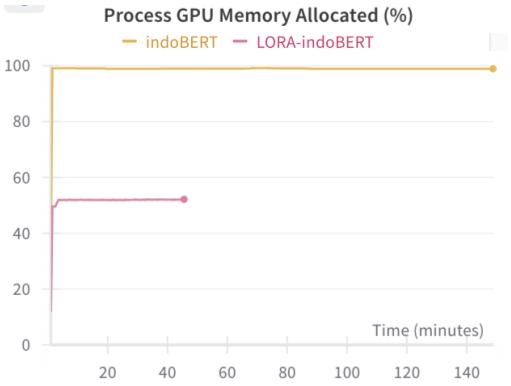


Figure 6. GPU memory utilization

4. CONCLUSION

This study has successfully demonstrated the effectiveness of utilizing pre-trained language models, especially CahyaBERT, IndoBERT, and MBERT, with full-parameter fine-tuning and LoRA-based tuning techniques for multi-label classification of diplomatic cables. This research finds that the Indonesian-specific model, particularly IndoBERT, outperforms the multilingual model in accurately categorizing diplomatic cables into politics, economics, protocol and consular services, socio-cultural affairs, and administrative matters. LoRA demonstrated significant advantages in GPU memory consumption by 48% and training time by 69.7% compared to full fine-tuning. However, by focusing LoRA adaptation on the self-attention layer, the study revealed several challenges, where the model's ability to capture complex relationships between labels was slightly compromised. LoRA showed a small drop in performance metrics compared to the full fine-tuning method, by 3.87% lower in subset accuracy and 9.99% higher in hamming loss. Future research could explore optimizing LoRA's application across different model layers or exploring hybrid find-tuning approaches that can better preserve semantic complexity. Ultimately, this experiment shows that LoRA is a promising approach for diplomatic institutions with limited computational resources, bridging the gap between model performance and computational efficiency.

FUNDING INFORMATION

No funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dela Nurlaila	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Abba Suganda Girsang		✓		✓	✓	✓				✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are restricted and not available to the public due to confidentiality requirements from the data provider institution.

REFERENCES

[1] United Nations, "Vienna convention on diplomatic relations (1961)," *Max Planck Encyclopedia of Public International Law*, 2009. [Online]. Available: https://treaties.un.org/doc/Treaties/1964/06/19640624 02-10 AM/Ch_III_3p.pdf. [Accessed Aug. 10, 2024].

[2] S. Barman, "Digital diplomacy: the influence of digital platforms on global diplomacy and foreign policy," *Vidya - a Journal of Gujarat University*, vol. 3, no. 1, pp. 61–75, 2024, doi: 10.47413/vidya.v3i1.304.

[3] C. Fernandes, "The cables and their reception," in *What uncle Sam Wants*, Palgrave Pivot Singapore, 2019.

[4] C. Bjola, J. Cassidy, and I. Manorc, "Public diplomacy in the digital age," *The Hague Journal of Diplomacy*, vol. 14, no. 1–2, pp. 83–101, 2019, doi: 10.1163/1871191X-14011032.

[5] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognition*, vol. 118, 2021, doi: 10.1016/j.patcog.2021.107965.

[6] V. S. Tidake and S. S. Sane, "Evaluation of multi-label classifiers in various domains using decision tree," *Advances in Intelligent Systems and Computing*, vol. 673, pp. 117–127, 2018, doi: 10.1007/978-981-10-7245-1_13.

[7] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 194–206, 2019, doi: 10.1007/978-3-030-32381-3_16.

[8] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, "Fine-tuning pretrained language models: weight initializations, data orders, and early stopping," *arXiv Computer Science*, 2020.

[9] E. Hu *et al.*, "LoRa: Low-rank adaptation of large language models," *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *Early Human Development*, vol. 83, no. 1, pp. 1–11, 2019, doi: 10.1016/j.earlhumdev.2006.05.022.

[11] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5641–5652, 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[12] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[13] E. T. Luthfi, Z. I. M. Yusoh, and B. M. Aboobaider, "BERT based named entity recognition for automated hadith narrator identification," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 604–611, 2022, doi: 10.14569/IJACSA.2022.0130173.

[14] K. C. Chien, C. H. Chang, and R. Der Sun, "Using parameter efficient fine-tuning on legal artificial intelligence," in *CEUR Workshop Proceedings*, vol. 3637, 2023.

[15] Ministry of Foreign Affairs, "Decree of the Minister of Foreign Affairs Number SK.06/A/OT/VI/2004/01 year 2004 concerning Organization and Working Procedures of the Republic of Indonesia's representatives abroad", (in Bahasa: Keputusan Menteri Luar Negeri Nomor SK.06/A/OT/VI/2004/01 Tahun 2004 tentang Organisasi dan Tata Kerja Perwakilan Republik Indonesia di Luar Negeri), 2004.

[16] A. Kurniasih and L. P. Manik, "On the role of text preprocessing in BERT embedding-based DNNs for classifying informal texts," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 927–934, 2022, doi: 10.14569/IJACSA.2022.01306109.

[17] I. O. Muraina, "Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts," in *7th International Mardin Artuklu Scientific Researches Conference*, pp. 496–504, 2022.
[18] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 145–158, 2011, doi: 10.1007/978-3-642-23808-6_10.
[19] J. Lee *et al.*, "K-MHaS: a multi-label hate speech detection dataset in Korean online news comment," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14264–14278, 2022, doi: 10.18653/v1/2023.findings-emnlp.952.
[20] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: a benchmark dataset and pre-trained language model for Indonesian NLP," *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.
[21] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 4996–5001, 2020, doi: 10.18653/v1/p19-1493.
[22] H. M. Zahera, I. Elgendy, R. Jalota, and M. A. Sherif, "Fine-tuned BERT model for multi-label tweets classification," in *28th Text REtrieval Conference, TREC 2019 - Proceedings*, pp. 1–7, 2019.
[23] M. Bilal and A. A. Almazroi, "Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews," *Electronic Commerce Research*, vol. 23, no. 4, pp. 2737–2757, 2023, doi: 10.1007/s10660-022-09560-w.
[24] Y. Mao *et al.*, "A survey on LoRA of large language models," *Frontiers of Computer Science*, vol. 19, no. 7, 2025, doi: 10.1007/s11704-024-40663-9.
[25] K. Dhole *et al.*, "NL-augmenter a framework for task-sensitive natural language augmentation," *Northern European Journal of Language Technology*, vol. 9, no. 1, 2023, doi: 10.3384/nejlt.2000-1533.2023.4725.
[26] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2018, pp. 489–500, doi: 10.18653/v1/d18-1045.
[27] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, vol. 1, pp. 86–96, 2016, doi: 10.18653/v1/p16-1009.
[28] M. Graça, Y. Kim, J. Schamper, S. Khadivi, and H. Ney, "Generalizing back-translation in neural machine translation," in *WMT 2019 - 4th Conference on Machine Translation, Proceedings of the Conference*, 2019, vol. 1, pp. 45–52, doi: 10.18653/v1/w19-5205.
[29] V. Ganganwar and R. Rajalakshmi, "MTDOT: A multilingual translation-based data augmentation technique for offensive content identification in Tamil text data," *Electronics*, vol. 11, no. 21, 2022, doi: 10.3390/electronics11213574.
[30] J. Van Nooten and W. Daelemans, "Improving dutch vaccine hesitancy monitoring via multi-label data augmentation with GPT-3.5," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 251–270, 2023, doi: 10.18653/v1/2023.wassa-1.23.
[31] D. Devi, S. K. Biswas, and B. Purkayastha, "A review on solution to class imbalance problem: undersampling approaches," in *2020 International Conference on Computational Performance Evaluation, ComPE 2020*, 2020, pp. 626–631, doi: 10.1109/ComPE49325.2020.9200087.

BIOGRAPHIES OF AUTHORS

Dela Nurlaila is a graduate student at Bina Nusantara University, pursuing a master's degree in computer science. She currently works at the Ministry of Foreign Affairs of the Republic of Indonesia. Her research interests include the application of artificial intelligence and data analytics in diplomatic communications and international relations. She can be contacted at email: dela.nurlaila@binus.ac.id.

Abba Suganda Girsang is currently a lecturer at Master of Computer Science, Bina Nusantara University since 2015. He got Ph.D. in the Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, he graduated bachelor from the Department of Electrical Engineering, Gadjah Mada University (UGM), Yogyakarta, Indonesia, in 2000. He then continued his master's degree in the Department of Computer Science in the same university in 2006–2008. He was a staff consultant programmer in Bethesda Hospital, Yogyakarta, in 2001 and worked as a web developer in 2002–2003. He then joined the Department of Informatics Engineering in Janabadra University as a lecturer in 2003-2015. He can be contacted at email: agirsang@binus.edu.