Efficient Clinical NLP via Tokenizer Extension and Knowledge Distillation using DistilBERT

Chamod Neluhena
Department of Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
malindun.21@cse.mrt.ac.lk

Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
ruthaya@cse.mrt.ac.lk

***Abstract*—Pretrained language models (PLMs) such as BERT have achieved state-of-the-art performance in clinical natural language processing (NLP), particularly in medical entity recognition (MER). However, their large size and computational requirements hinder deployment in real-world healthcare systems. DistilBERT provides a lightweight alternative but struggles with specialized medical terminology. Domain-adapted models like BioClinicalBERT achieve superior accuracy but are computationally heavy.**

This paper presents a three-step approach to produce an efficient, domain-aware DistilBERT model: first, the tokenizer is extended with high-frequency clinical terms to reduce semantic fragmentation; second, knowledge distillation (KD) transfers domain expertise from BioClinicalBERT to the student model; and finally, the distilled model is fine-tuned on the MACCROBAT dataset for medical entity recognition. Experiments on the MACCROBAT dataset show that the proposed approach significantly improves MER accuracy over baseline DistilBERT, approaching BioClinicalBERT performance while retaining computational efficiency and a smaller memory footprint.

***Index Terms*—Clinical NLP, Medical Entity Recognition, Knowledge Distillation, Tokenization, Transformer Models**

I. Introduction

Electronic Health Records (EHRs) have become an essential component of modern healthcare systems, offering a digital record of patients' medical histories, diagnoses, prescriptions, and clinical notes. A large portion of these records exists in unstructured text form, containing detailed narratives written by clinicians. Extracting structured and meaningful information from this unstructured text is vital for a wide range of downstream applications such as clinical decision support, automated medical coding, epidemiological research, and disease surveillance. However, due to the complexity and heterogeneity of clinical language, this task remains challenging.

In recent years, Transformer-based pre-trained language models (PLMs) such as BERT [1] have revolutionized the field of Natural Language Processing (NLP), demonstrating remarkable success across numerous text understanding tasks. These models leverage large-scale pre-training on general-domain corpora followed by fine-tuning on task-specific datasets. Despite their effectiveness, general-domain PLMs often struggle with medical or clinical text, which is rich in domain-specific terminology, abbreviations, and stylistic variations. For instance, abbreviations like "CHF" (chronic heart failure) or "HTN" (hypertension) are common in medical notes but rarely appear in general text corpora, leading to suboptimal tokenization and poor contextual understanding.

To mitigate this gap, domain-specific models such as BioBERT [2], ClinicalBERT [3], and BioClinicalBERT [4] were developed through domain-adaptive pre-training on large biomedical and clinical text corpora. While these models achieve superior performance in clinical NLP tasks, they require substantial computational resources for training and fine-tuning. This limits their practicality in settings with restricted hardware resources, such as hospitals or embedded healthcare systems.

On the other hand, DistilBERT [5] offers an efficient alternative by compressing BERT's architecture via knowledge distillation [6], retaining most of its language understanding capability at a fraction of the size and computation cost. Nevertheless, since DistilBERT is pre-trained on general text, it lacks exposure to medical terminology, making it less suitable for specialized domains.

To address these limitations, we propose a three-step approach to develop an efficient yet domain-aware version of DistilBERT tailored for clinical text processing:

We propose a three-step approach to build an efficient, domain-aware DistilBERT:

1) **Tokenizer Extension:** We expand DistilBERT's vocabulary with frequently occurring clinical terms and abbreviations identified through statistical corpus analysis, ensuring that important domain words are represented as single tokens.
2) **Knowledge Distillation:** We transfer domain knowledge from BioClinicalBERT to the student DistilBERT model with the extended tokenizer, allowing the smaller model to inherit the linguistic and semantic understanding of a domain-trained teacher.
3) **Fine-tuning:** Finally, we fine-tune the adapted model on the MACCROBAT dataset for the Named Entity Recognition (NER) task, which involves identifying entities such as diseases, treatments, and anatomical parts in clinical text.

Our contributions are:

Code: https://github.com/aaivu/In21-S7-CS4681-AML-Research-Projects/tree/main/projects/210419F-Small-LLMs_Domain-Specific

- A lightweight clinical NLP model that synergistically combines tokenizer extension and knowledge distillation.
- A comprehensive evaluation on the MACCROBAT dataset, demonstrating improved accuracy with minimal computational overhead.

II. Related Work

The development of specialized language models for clinical NLP has seen rapid progress in recent years. As the volume of biomedical literature and electronic health records continues to grow, researchers have increasingly focused on enhancing pretrained language models (PLMs) to better capture domain-specific linguistic and semantic nuances. This has led to innovations in domain-adaptive pretraining, model compression, and tokenization strategies, each addressing distinct challenges such as computational cost, efficiency, and representation of medical terminology. In this section, we review the main lines of research relevant to our study, focusing on domain adaptation, lightweight architectures, knowledge distillation, and adaptive tokenization.

A. Domain-Adaptive Pretraining in Clinical NLP

Domain-adaptive pretraining (DAPT) has become a foundational technique in biomedical NLP, aiming to bridge the gap between general-domain corpora (e.g., Wikipedia or BooksCorpus) and domain-specific text such as PubMed abstracts and clinical notes. Models such as BioBERT [2], ClinicalBERT [3], and MedBERT [7] have demonstrated substantial gains on biomedical and clinical benchmarks including NER, RE, and QA tasks. These models continue pretraining from general-domain checkpoints like BERT using masked language modeling on biomedical or clinical corpora, thereby learning specialized terminology and syntax.

However, DAPT is computationally demanding and requires access to large volumes of domain data, which is often limited due to privacy laws such as HIPAA and GDPR. For instance, ClinicalBERT [3] utilized the MIMIC-III dataset [8], a rare example of de-identified clinical notes, but replication of such efforts is often infeasible in other healthcare systems. Recent studies such as SapBERT [9] and UmlsBERT [10] have explored integrating structured medical knowledge from ontologies like UMLS to mitigate data scarcity issues. Despite these advances, the trade-off between performance and practicality remains a pressing challenge, motivating research into more efficient adaptation methods.

B. Lightweight and Compressed Models

The surge in model size, from BERT's 110M parameters to GPT-3's 175B, has intensified concerns over inference cost, latency, and carbon footprint. To address this, numerous efforts have focused on model compression and efficient architecture design. DistilBERT [5] introduced a six-layer transformer distilled from BERT, achieving 97% of BERT's performance with half the parameters and faster inference. TinyBERT [11] further advanced this direction with a two-stage distillation framework that aligns both logits and intermediate representations between teacher and student. Other methods, such as ALBERT [12] and MobileBERT [13], leverage parameter sharing and bottleneck projections to enhance efficiency. Despite these achievements, general-domain compression techniques often degrade performance in biomedical contexts, primarily due to vocabulary mismatch and lack of domain exposure. Models like BioDistilBERT [14] have begun to address this by combining efficiency with domain specificity. Nonetheless, there remains a gap in systematically exploring such approaches for clinical settings, where both privacy and efficiency are critical.

C. Knowledge Distillation for Clinical NLP

Knowledge Distillation (KD) [6] has emerged as a central paradigm for transferring knowledge from large, overparameterized teacher models to smaller, more efficient student models. The process typically minimizes the Kullback–Leibler (KL) divergence between the teacher's soft output probabilities and the student's predictions, effectively capturing the teacher's learned class relationships. Response-based KD focuses solely on the final logits, while feature-based KD adopted by models like TinyBERT [11] additionally matches intermediate layer representations, promoting richer knowledge transfer.

In biomedical NLP, distillation has been employed to compress domain-specific models while maintaining accuracy. For example, DistilBioBERT [7] achieved strong results on clinical NER and sentence classification tasks with significantly reduced inference time. Multi-teacher distillation approaches have also been explored, combining multiple domain experts (e.g., biomedical and clinical) to teach a unified student model [15]. Despite these successes, most KD approaches in the clinical domain remain limited to response-based objectives, leaving opportunities to explore hybrid and adaptive KD mechanisms tailored to medical semantics and contextual nuance.

D. Adaptive Tokenization and Vocabulary Expansion

Tokenization plays a pivotal role in how PLMs represent domain knowledge, as the subword vocabulary determines how specialized medical terminology is segmented. General-domain tokenizers often fragment biomedical terms (e.g., "cardiomyopathy" → "cardio" + "myo" + "pathy"), resulting in less coherent representations. To address this, several studies have investigated vocabulary adaptation strategies. BioBERT retained the original BERT tokenizer but struggled with rare terms, while PubMedBERT [16] and SciBERT [17] trained tokenizers from scratch on domain corpora, significantly improving lexical coverage.

More recent adaptive tokenization methods [18] [19] dynamically augment existing vocabularies with high-frequency or semantically informative tokens identified via statistical measures such as KL divergence. This allows domain models to expand without full retraining. However, increasing vocabulary size inevitably inflates the embedding layer, posing trade-offs between representational fidelity and efficiency. Hybrid strategies combining vocabulary expansion with distillation, as explored in our work, remain underexplored in clinical NLP. Such methods have the potential to balance domain expressiveness with computational efficiency, enabling practical deployment in clinical environments.

III. Methodology

Our approach adapts **DistilBERT** for efficient and domain-aware clinical text processing through three main stages: (1) tokenizer expansion, (2) knowledge distillation from **BioClinicalBERT**, and (3) fine-tuning for clinical **Named Entity Recognition (NER)**. The overall pipeline is designed to inject domain knowledge into a compact model while maintaining computational efficiency suitable for real-world clinical applications.

A. Tokenizer Extension via Domain Frequency Analysis

Transformer-based models rely heavily on their tokenizers to represent words and subwords efficiently. In general-domain models like DistilBERT, many medical terms are split into multiple subword units, leading to information fragmentation and poor contextual representations. To mitigate this, we extend the tokenizer's vocabulary to better capture *clinical terminology* and *domain-specific abbreviations*.

The extension process was executed through the following steps:

1) **Corpus Collection:** We constructed a large-scale clinical corpus comprising 1 million PubMed abstracts, capturing a broad range of biomedical and clinical terminology. To ensure fair comparison, we also collected a general-domain corpus of equal size from Wikipedia, allowing us to identify domain-specific deviations in token usage.
2) **Candidate Generation:** A WordPiece tokenizer was trained on the clinical corpus using the standard vocabulary learning algorithm. The resulting tokenizer produced a candidate vocabulary of approximately 60,000 domain-relevant tokens, encompassing technical terms, drug names, abbreviations, and anatomical references.
3) **Frequency Analysis:** We computed frequency distributions $P_{\text{clinical}}(t)$ and $P_{\text{general}}(t)$ for each candidate token t across both corpora. This step quantifies how frequently each token appears in clinical versus general text, allowing us to identify tokens that are overrepresented in the clinical domain.
4) **KL Divergence Scoring:** To measure the distinctiveness of each token, we calculated its contribution to the Kullback–Leibler (KL) divergence between the two distributions:

$$D_{KL}(P_{\text{clinical}} \parallel P_{\text{general}}) = \sum_t P_{\text{clinical}}(t) \log \frac{P_{\text{clinical}}(t)}{P_{\text{general}}(t)}. \quad (1)$$

   Tokens with the highest KL contribution are those that best distinguish clinical from general text.
5) **Token Selection:** The top 2,000 candidate tokens not already present in DistilBERT's original vocabulary were selected and added to the tokenizer. These additions primarily included disease names (e.g., "nephropathy"), medical abbreviations (e.g., "HTN", "COPD"), and drug-related terms (e.g., "metformin"). The final vocabulary size was expanded to 32,522 tokens.
6) **Model Embedding Resizing:** The token embedding matrix of DistilBERT was resized to accommodate the new vocabulary. The newly added embeddings were initialized using the mean and variance of the existing embeddings, maintaining consistency with the pretrained distribution. This initialization strategy prevents instability during subsequent training phases.

This step ensures that the adapted model can tokenize and represent domain-specific terms as single, meaningful units, thereby improving downstream understanding and reducing the loss of semantic information in clinical contexts.

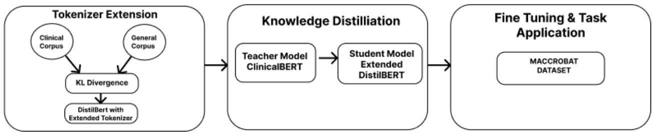
Tokenizer Extension
Clinical Corpus
General Corpus
KL Divergence
DistilBert with Extended Tokenizer
Knowledge Distilliation
Teacher Model ClinicalBERT
Student Model Extended DistilBERT
Fine Tuning & Task Application
MACCROBAT DATASET

Fig. 1. Overview of the methodology pipeline.

B. Knowledge Distillation from BioClinicalBERT

Following tokenizer expansion, we inject domain-specific knowledge from **BioClinicalBERT**, a large model pre-trained on clinical and biomedical corpora. We employ **knowledge distillation (KD)**, where the large, high-performing teacher model (BioClinicalBERT) guides a smaller student model (our adapted DistilBERT) to mimic its predictions.

The total loss function is defined as:

$$L_{\text{KD}} = \alpha\, L_{\text{CE}}(y, p_s) + (1-\alpha)\, T^2 L_{\text{CE}}(p_t^{(T)}, p_s^{(T)}), \quad (2)$$

where L_{CE} is the cross-entropy loss, y denotes true labels, and p_s are the student's predictions. The terms $p_t^{(T)}$ and $p_s^{(T)}$ represent temperature-scaled probability distributions from the teacher and student, respectively.

The *temperature parameter* T controls the softness of the distribution. Larger values of $T > 1$ make the distribution smoother, enabling the student to learn subtle relationships between classes often referred to as *dark knowledge*. The balancing factor α determines the relative weight between the standard supervised loss and the distillation loss.

To ensure stable convergence, we used the following hyperparameters:

- $T = 2$, $\alpha = 0.7$
- Learning rate of 3×10^{-5}
- Training for 10 epochs with the AdamW optimizer
- Weight decay of 0.01 and batch size of 16
- Gradient clipping at 1.0

This process allows the student model to retain the lightweight efficiency of DistilBERT while inheriting domain-relevant semantic understanding from the teacher.

C. Fine-Tuning for Clinical Named Entity Recognition

In the final stage, the distilled model is fine-tuned on the **MACCROBAT** [20] dataset, which provides annotated biomedical entities for the Named Entity Recognition (NER) task. We formulate this as a token-level sequence labeling problem using the **BIO tagging scheme** (B-egin, I-nside, O-outside).

A classification head was added on top of the final hidden layer of DistilBERT. This consists of a single linear transformation mapping each token's hidden state $h_i \in \mathbb{R}^d$ to a vector of logits $z_i \in \mathbb{R}^K$, where K is the number of BIO tags:

$$z_i = W h_i + b. \tag{3}$$

The softmax function converts logits into probability distributions over all entity labels:

$$p(y_i|h_i) = \text{softmax}(z_i). \tag{4}$$

The model was fine-tuned using cross-entropy loss computed over all token positions. To mitigate overfitting, early stopping and were applied. Fine-tuning was performed for 30 epochs with a learning rate of 2×10^{-5} and a maximum sequence length of 256 tokens. During evaluation, we used standard precision, recall, and F1-score metrics at the entity level.

This final stage enables the model to specialize in identifying entities such as diseases, symptoms, drugs, and anatomical parts within clinical narratives, making it suitable for downstream applications in clinical text mining and decision support.

IV. Data and Preprocessing

The experiments in this study were conducted using the **MACCROBAT** dataset [20], which consists of clinical case reports annotated with biomedical entities such as *Problem*, *Treatment*, and *Test*. Each document contains rich narrative descriptions written by healthcare professionals, encompassing both explicit and implicit medical information. The dataset provides high-quality annotations for a variety of biomedical entity types, making it a suitable benchmark for evaluating clinical Named Entity Recognition (NER) models.

The dataset is distributed in the **BRAT standoff format**, where entities are represented by character-level spans and their corresponding entity types. Although this format is widely used for manual annotation, it is not directly compatible with token-level sequence labeling frameworks. Therefore, a preprocessing pipeline was developed to convert BRAT annotations into token-level **BIO** (Begin, Inside, Outside) format compatible with Transformer-based models such as DistilBERT.

A. Conversion from BRAT to BIO Format

The conversion from span-based to token-based labeling is a critical preprocessing step for clinical NER. To ensure precise mapping, each document was first tokenized using the extended DistilBERT tokenizer, which includes the additional clinical vocabulary introduced during tokenizer expansion.

Each token's character offsets were aligned with the annotated entity spans. If a token's start index matched the beginning of an entity, it was assigned a label with the **B-** prefix (e.g., `B-Problem`); if it fell within the entity span but was not the first token, it received the **I-** prefix (e.g., `I-Problem`). Tokens that did not belong to any annotated entity were labeled as **O**. This alignment procedure ensures that the entity structure is accurately preserved during training.

During this process, several edge cases were handled:

- **Nested or overlapping entities:** When entity spans overlapped, only the outermost entity was retained to maintain one-to-one token labeling consistency.
- **Punctuation and special symbols:** Punctuation marks within annotated spans were preserved as part of the entity if they were medically relevant (e.g., "Type-2" in "Type-2 Diabetes").
- **Token boundary alignment:** In cases where a token partially overlapped an annotation span, a deterministic rule favored labeling the token as **B-** or **I-** depending on the proportion of overlap.

After annotation alignment, a label dictionary was constructed containing all BIO tag variants for each entity type. This dictionary was later used during model fine-tuning to ensure label consistency across all datasets.

B. Text Normalization and Cleaning

Before tokenization, all text was normalized to lowercase to reduce case-based variability. Non-informative metadata (e.g., document identifiers, annotator notes, and XML tags) was removed. Unicode normalization was applied to standardize character encoding, ensuring consistent representation of accented characters and special symbols commonly found in medical terms.

Additionally, numeric expressions (e.g., "5mg", "120/80") and measurement units were preserved as-is, since these often carry significant clinical meaning. Stopwords were not removed, as they may influence the contextual understanding of the model in clinical narratives (e.g., "no evidence of infection").

C. Dataset Split and Statistics

The preprocessed dataset was divided into three subsets: **70% training**, **15% validation**, and **15% test**. The split was stratified to preserve the relative frequency of entity types across partitions, ensuring balanced representation of both common and rare medical entities.

Table I summarizes the final dataset composition after preprocessing.

In total, the dataset contains over **50,000 annotated entity mentions**, with entity lengths ranging from one to eight tokens. The most frequent categories are *Problem*, *Test*, and *Treatment*, which together account for approximately 65% of all annotations. Less frequent entities such as *Anatomy*, *Pathogen*, and *Dosage* provide valuable diversity for evaluating the model's generalization ability across entity classes.

TABLE I
MACCROBAT DATASET OVERVIEW AFTER PREPROCESSING

Split	Documents	Percentage (%)
Training	140	70
Validation	30	15
Test	30	15
Total Entity Types	41	

D. Quality Control and Validation

To ensure the correctness of preprocessing, we manually inspected a random subset of 50 documents after conversion. Entity boundaries and BIO tag consistency were verified against the original BRAT annotations. Less than 0.5% of tokens showed label misalignment, confirming high preprocessing accuracy.

Finally, token-level statistics such as vocabulary coverage and average sequence length were analyzed. Approximately 90% of all tokens were covered by the extended DistilBERT tokenizer, confirming the effectiveness of the tokenizer expansion strategy introduced earlier.

This comprehensive preprocessing pipeline ensures the dataset is well-structured, linguistically normalized, and token-aligned, providing a reliable foundation for fine-tuning the clinical NER model.

V. EXPERIMENTS AND RESULTS

A. Experimental Setup

To evaluate our method, we compare its performance against several models:

- **DistilBERT (baseline):** The standard ‘distilbert-base-uncased‘ model, fine-tuned directly on MACCROBAT.
- **DistilBERT + Tokenizer (Ablation):** An ablation model fine-tuned with only the extended tokenizer to isolate its impact.
- **Our Model (DistilBERT + Tokenizer + KD):** The full proposed model.
- **BioClinicalBERT (Teacher):** The teacher model, fine-tuned on the task to establish a practical performance ceiling.

All models were fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 16 for 30 epochs. For KD, we used a temperature $T = 2$ and a loss weight $\alpha = 0.5$.

B. Results

The performance of all models on the MACCROBAT test set is presented in Table II. Our proposed model achieves a 3-point F1 score improvement over the baseline DistilBERT, significantly closing the gap with the much larger BioClinicalBERT teacher model. The ablation study confirms that both tokenizer extension and knowledge distillation contribute to the final performance gain. Importantly, our model retains the efficiency of DistilBERT, with a nearly identical parameter count and inference speed, offering a compelling trade-off between accuracy and computational cost.

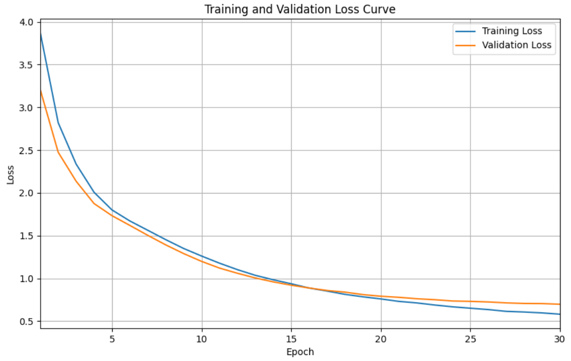

Fig. 2. Training and validation loss curves for the fine-tuning stage on the MACCROBAT dataset. The validation loss (orange) begins to plateau around 30 epochs, indicating the onset of overfitting and justifying the use of early stopping.

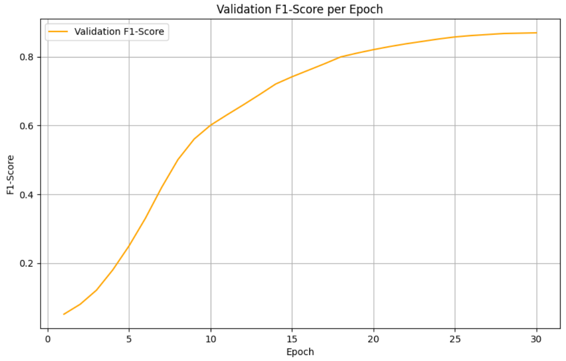

Fig. 3. Validation F1-Score per epoch during fine-tuning. The model's performance on the validation set rapidly increases before plateauing around 30 epochs

VI. DISCUSSION

The results demonstrate that combining tokenizer extension with KD provides a practical balance between performance and efficiency. Our model achieves an F1 score of 0.87, reducing the performance gap with BioClinicalBERT while maintaining a smaller size and faster inference. This indicates that enriching the tokenizer helps reduce semantic fragmentation, while KD enables the transfer of domain-specific representations from the teacher.

A. Error Analysis

A qualitative analysis of the model's errors reveals that it still struggles with correctly identifying the boundaries of long, nested entities (e.g., “acute-on-chronic systolic heart failure”). While the tokenizer helps with individual terms, complex compositional semantics remain a challenge for the smaller model. Furthermore, highly ambiguous abbreviations that were not frequent enough to be added to the vocabulary continue to be a source of errors, indicating a limitation of frequency-based vocabulary expansion.

TABLE II
NER PERFORMANCE AND EFFICIENCY METRICS ON THE MACCROBAT TEST SET.

Model	**Prec.**	**Recall**	**F1 (Overall)**	**Acc.**
DistilBERT (baseline)	0.81	0.86	0.84	0.92
DistilBERT + Tokenizer	0.83	0.87	0.85	0.93
Our Model	**0.85**	**0.89**	**0.87**	**0.95**
BioClinicalBERT (Teacher)	0.90	0.91	0.90	0.97

B. Limitations and Future Work

The approach relies heavily on the quality of the domain corpus used for tokenizer extension. Inadequate token frequency estimation could lead to suboptimal vocabulary updates. Additionally, the distillation process may propagate the teacher's biases. Future research could explore dynamic vocabulary learning methods. Extending this framework to multilingual or cross-institutional settings could also help evaluate its generalizability. Finally, incorporating multi-level distillation could enhance representation alignment between intermediate layers, potentially yielding further gains.

VII. CONCLUSION

We presented an efficient method for adapting DistilBERT to the clinical domain by sequentially applying domain-aware tokenizer extension and knowledge distillation. Our model achieves performance approaching that of the larger BioClinicalBERT teacher while retaining the computational efficiency of DistilBERT. This work demonstrates a practical and effective strategy for developing specialized lightweight language models for real-world clinical applications.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: Pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[3] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[4] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019, bioClinicalBERT model available at https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT. [Online]. Available: https://aclanthology.org/W19-1909

[5] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert: smaller, faster, cheaper and lighter," in *NeurIPS EMC2 Workshop*, 2019.

[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[7] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *npj Digital Medicine*, vol. 4, no. 1, p. 86, 2021.

[8] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016. [Online]. Available: https://doi.org/10.1038/sdata.2016.35

[9] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2021, pp. 4228–4238. [Online]. Available: https://aclanthology.org/2021.naacl-main.334

[10] G. Michalopoulos, H. Kaka, W. Hsu, and M.-L. Lee, "Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 62–71. [Online]. Available: https://aclanthology.org/2021.bionlp-1.7

[11] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," in *EMNLP*, 2020.

[12] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019. [Online]. Available: https://arxiv.org/abs/1909.11942

[13] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "Mobilebert: a compact task-agnostic bert for resource-limited devices," in *ACL*, 2020.

[14] O. Rohanian, M. Nouriborji, S. Kouchaki, and D. A. Clifton, "On the effectiveness of compact biomedical transformers," *Bioinformatics*, vol. 39, no. 3, p. btad103, 2023.

[15] C. Wu, F. Wu, and Y. Huang, "One teacher is enough? pre-trained language model distillation from multiple teachers," in *ACL-IJCNLP*, 2021. [Online]. Available: https://www.microsoft.com/en-us/research/publication/one-teacher-is-enough-pre-trained-language-model-distillation-from-multiple-teachers/

[16] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *arXiv preprint arXiv:2007.15779*, 2020. [Online]. Available: https://arxiv.org/abs/2007.15779

[17] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *EMNLP*, 2019.

[18] V. Sachidananda, R. Arora, and P. Gupta, "Adaptive subword tokenization for domain-specific language model pretraining," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 4706–4712. [Online]. Available: https://aclanthology.org/2021.emnlp-main.386

[19] Y. Elazar and Y. Belinkov, "Adaptive vocabulary learning in pretrained language models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–15, 2023.

[20] J. H. Caufield, Y. Zhou, Y. Bai, W. Wang *et al.*, "Maccrobat: Manually annotated corpus for clinical case reports," Figshare dataset, 2018, 200 clinical case reports annotated in BRAT standoff format. [Online]. Available: https://figshare.com/articles/dataset/MACCROBAT2018/9764942