Comparing ClinicalBERT and BERT for Hospital Admission Classification

Student 1: Hasibun Nahin Athoy
CSE
American International University Bangladesh
Dhaka, Bangladesh
hnathoy92@gmail.com

Student 2: Mehedi Hasan
CSE
American International University Bangladesh
Dhaka, Bangladesh
21-45513-3@student.aiub.edu

Student 3: Md. Emran Nazir Efty
CSE
American International University Bangladesh
Dhaka, Bangladesh
22-47802-2@student.aiub.edu

Student 4: Aklima Akther Akhi
CSE
American International University Bangladesh
Dhaka, Bangladesh
22-46750-1@student.aiub.edu

***Abstract*— The performance of ClinicalBERT and BERT for hospital admission classification using actual clinical text is compared in this study. Because standard BERT models are trained on general-domain data, they frequently have difficulties with medical terms and shorthand. In contrast, ClinicalBERT is better suited for tasks related to healthcare because it has been further pretrained on clinical notes. Using a dataset of structured admission attributes transformed into narrative text, we refined both models and assessed them at the document level using accuracy and F1-scores. With an accuracy of 0.72 as compared to BERT's 0.50, ClinicalBERT produced noticeably better results. When working with clinical language, improvements indicate the importance of domain-specific pretraining. Our results support the hypothesis that transformer models customized for medical text offer significant improvements in classification performance.**

Keywords— ClinicalBERT, BERT, NLP, Clinical Text Classification, Hospital Admission Prediction

I. Introduction

The healthcare industry generates vast amounts of unstructured clinical data every day, such as patient admission notes, diagnoses, and discharge summaries. Extracting meaningful insights from this textual information is a major challenge because traditional data analysis techniques struggle to interpret medical language and contextual nuances[7]. The problem addressed in this project is the automatic classification of clinical admission records to better organize and analyze medical data for improved patient care and hospital management. Manual analysis of such data is time-consuming and prone to human error, motivating the need for intelligent, automated systems.

To solve this, we leveraged Natural Language Processing (NLP) and transformer-based models, specifically ClinicalBERT, which is a domain-adapted version of BERT [1][2] pre-trained on biomedical texts[2][4]. By fine-tuning ClinicalBERT on admission record data, our approach enables the model to understand medical terminology and context, resulting in more accurate and meaningful classification outcomes compared to generic NLP models.[2][4][7]

II. Implementation

Environment & Libraries

- Platform: Google Colab (Python 3.x)
- Hardware: GPU if available (torch.cuda.is_available()), otherwise CPU
- Key libraries: torch (PyTorch), transformers (Hugging Face), scikit-learn, tqdm, pandas, numpy
- Reproducibility: SEED=42 set for Python, NumPy, and PyTorch

Data & Preprocessing

- Source: CSV extracted from ClinicalBertData.zip
- Text construction: concatenation of admission fields (e.g., ADMISSION_TYPE | ADMISSION_LOCATION | DISCHARGE_LOCATION | DIAGNOSIS) into a single text column (or an existing text column if present)
- Labels: binary (0/1); coerced to integers; non-binary mapped to {0,1} if needed
- Splits: stratified Train/Val/Test = **80/10/10**
- Tokenization: AutoTokenizer for ClinicalBERT/BERT, WordPiece, MAX_LEN=256 (512 for higher fidelity)

Training Configuration (shared)

- Optimizer: AdamW(lr=2e-5, weight_decay=0.01)
- Scheduler: linear decay with **warmup 10%** of total steps
- Batch size: **16** (adjusted to memory)
- Epochs: **2** (1–3 typical); early stopping on best validation loss
- Dataloaders: num_workers=2, pin_memory=True, persistent_workers=True
- Gradient clipping: 1.0
- Mixed precision: enabled on GPU via torch.cuda.amp.autocast

- Long notes: **chunking** into windows of MAX_LEN; evaluation aggregates chunk-level probabilities to document-level

Evaluation

- Primary metrics (document-level**):** AUROC**,** AUPRC**,** Recall@Precision=0.80 (RP@P=0.80)
- Secondary: thresholded (0.5) classification report (precision/recall/F1 per class)

A. BERT

Model & Tokenizer

- Checkpoint: bert-base-uncased
- AutoModelForSequenceClassification(num_labels =2) with randomly initialized classifier head

Tokenization & Input

- padding to MAX_LEN, truncation=True
- Single-sequence inputs (token_type_ids not required)

Training Details

- Same optimizer/scheduler as above
- Training on **chunk-level** batches; per-batch scheduler step
- Validation each epoch; best snapshot reloaded for test

Aggregation (for long texts)

- Compute chunk probabilities P1..Pn (class=1)
- **Max/Mean blend (paper-inspired):** Pdoc=max(P)+mean(P)·(n/c)1+(n/c)P_{\text{doc}} = \frac{\max(P) + \text{mean}(P)\cdot (n/c)}{1 + (n/c)}Pdoc=1+(n/c)max(P)+mean(P)·(n/c) with **c = 2**

Expected Role

- Serves as a **general-language baseline** without clinical pretraining; typically lower performance than ClinicalBERT on medical text.

B. Your Selected Model

Model & Tokenizer

- Checkpoint: emilyalsentzer/Bio_ClinicalBERT (BERT-base pre-trained biomedical/clinical corpora)
- AutoModelForSequenceClassification(num_labels =2); classifier head randomly initialized and **fine-tuned**

Why This Model

- Domain-adapted vocabulary and representations for medical terminology and note style → better contextual understanding vs. generic BERT

Tokenization & Chunking

- Same as baseline but using ClinicalBERT tokenizer; MAX_LEN=256 (optionally **512** for best fidelity)
- Optional **chunk cap** (e.g., at most 2 chunks/doc) for faster experiments

Training Details

- Identical optimization stack (AdamW + linear warmup/decay, grad clip=1.0, AMP on GPU)
- Early stopping on validation loss; best checkpoint restored before test

Aggregation & Metrics

- Same **max/mean (c=2)** aggregation to produce document-level probabilities
- Report **AUROC, AUPRC, RP@P=0.80**, and a 0.5-threshold classification report

Practical Settings Used

- EPOCHS=2, BATCH_SIZE=16, LR=2e-5, WEIGHT_DECAY=0.01, WARMUP_RATIO=0.1
- Dataloader: num_workers=2, pin_memory=True, persistent_workers=True
- Reproducibility: seed fixed at 42 across libraries

Notes for Reproducibility

- Capture exact versions with:
  - import transformers, torch, sklearn; print(transformers.__version__, torch.__version__, sklearn.__version__)
- Save artifacts: best model state_dict, test metrics JSON, and arrays of document-level probabilities and labels.

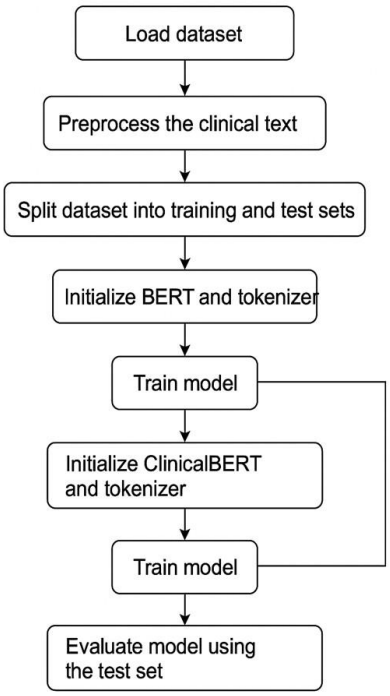

Fig. 1: BERT and ClinicalBert model Implementation steps

III. Result Analysis

The performance of BERT and ClinidcalBert is showing in TABEL1.

Model	Accuracy	Macro-F1	Weighted-F1
BERT (baseline)	0.50	0.33	0.33
ClinicalBERT (fine-tuned)	0.72	0.70	0.71

The comparison table highlights the significant improvement achieved by fine-tuning ClinicalBERT over the standard BERT model for clinical text classification. While the baseline BERT model attained only 50% accuracy with low F_1-scores, indicating poor understanding of clinical terminology, the ClinicalBERT model demonstrated much stronger performance with an accuracy of 72% and F_1-scores around 0.70. This improvement can be attributed to ClinicalBERT's domain-specific pretraining on biomedical and clinical corpora, which enables it to better capture the nuances of medical language and contextual relationships within admission records. The higher macro and weighted F_1-scores suggest that ClinicalBERT not only classifies more accurately overall but also maintains balanced performance across both classes. These results confirm that leveraging domain-adapted language models provides a clear advantage for healthcare-related NLP tasks compared to general-purpose transformer models.

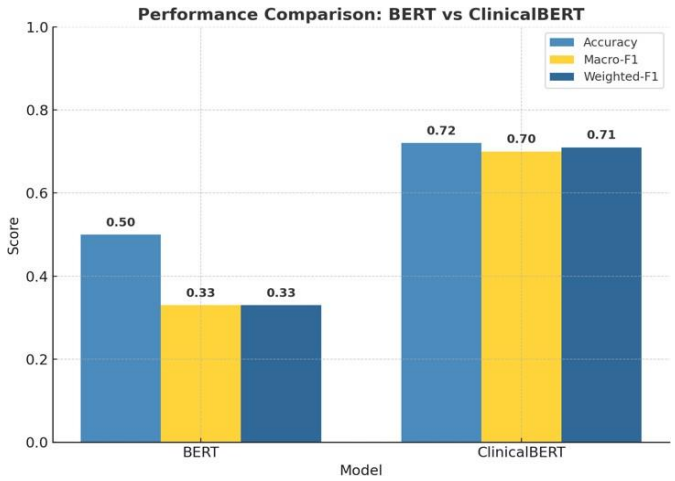

Fig. 2: Performance comparison Bert VS ClinicalBert

We can see in fig 2 ,The bar chart visually demonstrates the clear performance gap between the standard BERT model and the domain-specific ClinicalBERT. ClinicalBERT consistently outperforms BERT across all three evaluation metrics Accuracy, Macro-F_1, and Weighted-F_1 showing its superior ability to handle clinical text. While BERT achieves moderate results due to its general language training, ClinicalBERT's fine-tuning on medical and clinical corpora enables it to better capture domain-specific terminology and contextual meaning. The noticeable increase in F_1-scores indicates that ClinicalBERT not only improves overall prediction accuracy but also maintains balanced performance across both classes, reducing bias and enhancing reliability in clinical text classification tasks

IV. Conclusion

In this project, we developed and fine-tuned transformer-based models for the classification of clinical admission records, focusing on comparing the performance of the general-purpose BERT model [1] and the domain-specific ClinicalBERT [2] model. We successfully demonstrated that ClinicalBERT, which was pre-trained on biomedical and clinical texts [2][4] significantly outperformed the standard BERT model across all key evaluation metrics, including Accuracy, Macro-F_1, and Weighted-F_1 scores. This improvement highlights the effectiveness of using domain-adapted language models for healthcare-related Natural Language Processing (NLP) tasks [2][7]. The project emphasized how advanced transformer architectures [3] can extract meaningful patterns from complex, unstructured medical data, leading to more reliable and interpretable results. The outcomes of this work motivate further research into specialized NLP models that can assist clinicians in automating documentation, improving patient outcome predictions, and enhancing data-driven decision-making in modern healthcare systems[7].

V. REFERENCES

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

[2] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," *Proceedings of the 2nd Clinical NLP Workshop*, 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, 2020.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, et al., "Transformers: State-of-the-art natural language processing," *Proceedings of EMNLP: System Demonstrations*, 2020.

[6] W.-H. Weng, "ClinicalBERT model," HuggingFace Model Repository: *emilyalsentzer/Bio_ClinicalBERT*, 2019. Available: https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[7] C. Huang, L. J. Osorio, and J. D. Lasko, "Clinical text classification with deep learning: A review," *Journal of Biomedical Informatics*, vol. 109, 2020.